Generative Emotion Cause Explanation in Multimodal Conversations

Lin Wang
q3222345200@gmail.com
Northeastern University
Shenyang, Liaoning, China

Xiaocui Yang
yangxiaocui@cse.neu.edu.cn
Northeastern University
Shenyang, Liaoning, China

Shi Feng*
fengshi@cse.neu.edu.cn
Northeastern University
Shenyang, Liaoning, China

Daling Wang
wangdaling@cse.neu.edu.cn
Northeastern University
Shenyang, Liaoning, China

Yifei Zhang
zhangyifei@cse.neu.edu.cn
Northeastern University
Shenyang, Liaoning, China

Zhitao Zhang
windowschuang@163.com
Shenyang Women's and Children's Hospital
Shenyang, Liaoning, China

arXiv:2411.02430v3 [cs.CL] 4 Jun 2025

Abstract

Multimodal conversation, a crucial form of human communication, carries rich emotional content, making the exploration of the causes of emotions within it a research endeavor of significant importance. However, existing research on the causes of emotions typically employs an utterance selection method within a single textual modality to locate causal utterances. This approach remains limited to coarse-grained assessments, lacks nuanced explanations of emotional causation, and demonstrates inadequate capability in identifying multimodal emotional triggers. Therefore, we introduce a task—**Multimodal Emotion Cause Explanation in Conversation (MECEC)**. This task aims to generate a summary based on the multimodal context of conversations, clearly and intuitively describing the reasons that trigger a given emotion. To adapt to this task, we develop a new dataset (ECEM) based on the MELD dataset. ECEM combines video clips with detailed explanations of character emotions, helping to explore the causal factors behind emotional expression in multimodal conversations. A novel approach, FAME-Net, is further proposed, that harnesses the power of Large Language Models (LLMs) to analyze visual data and accurately interpret the emotions conveyed through facial expressions in videos. By exploiting the contagion effect of facial emotions, FAME-Net effectively captures the emotional causes of individuals engaged in conversations. Our experimental results on the newly constructed dataset show that FAME-Net outperforms several excellent baselines. Code and dataset are available at https://github.com/3222345200/FAME-Net.

CCS Concepts

• **Information systems** → **Sentiment analysis**; • **Computing methodologies** → *Natural language processing*.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMR '25, Chicago, IL, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733348

Keywords

Multimodal Emotion Cause Generation; Emotion Cause Analysis in Conversations; Large Language Model

ACM Reference Format:
Lin Wang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Zhitao Zhang. 2025. Generative Emotion Cause Explanation in Multimodal Conversations. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3731715.3733348

1 Introduction

Multimodal conversation plays a significant role in our daily life, influencing our thoughts and behaviors. Research [1, 13, 60] demonstrates that exploring the causes of emotions holds great importance, as a deeper understanding of human emotions and their triggers promotes mental health, optimizes interpersonal relationships, and fosters the development of more empathetic and humanized artificial intelligence systems [16, 18, 36]. In recent years, researchers have increasingly focused on how to explore the underlying causes of emotions. Consequently, a new task called Emotion Cause Extraction (ECE) [63, 70] is proposed, along with the introduction of a benchmark dataset named RECCON [48]. The goal of the ECE task is to identify the specific utterances in a conversation that causally trigger the emotion expressed in the target utterance.

Wang et al. [57] further propose a new task, Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE), which expands the research scope to include multimodal content. However, the identification of emotional causes in this task still relies on selective cause extraction rather than cause generation. Selective cause extraction can only perform a simple selection of emotional causes at a coarse-grained level and fails to capture fine-grained emotional causes that require deeper contextual reasoning. For those emotional utterances, if the causal relationships behind them are primarily conveyed through visual or auditory information, determining the specific causes that trigger emotions becomes particularly challenging. Therefore, generating detailed explanations for emotional causes can provide a more comprehensive and nuanced account of the underlying reasons behind emotions. The comparison between selective and generative approaches is illustrated in Figure 1. Wang et al. [58] introduce the Multimodal Emotion Cause Generation in Conversations (MECGC) task and the ECGF dataset to address



Figure 1: Comparison between selective and generative: (a) Selective task requires identifying causal statements (#7) in the target utterance (#7) that trigger the speaker's emotions. (b) Generative tasks require extracting the causes for the emotions contained in the target utterances from the conversation history.

limitations in selective emotion cause discovery. However, with an average explanation length of only 8.64 words, ECGF struggles to capture the complexity of emotion causes, potentially missing crucial contextual details and subtle emotional triggers.

To enable more comprehensive emotion cause analysis, we propose the **Multimodal Emotion Cause Explanation in Conversation (MECEC)** task. MECEC aims to generate detailed explanations for emotional expressions by leveraging both textual and visual information from conversations. We introduce the **Emotion Cause Explanation (ECEM)** dataset, built upon MELD [49], comprising 7,273 video clips with rich multimodal annotations. ECEM features an average explanation length of 14.4 words, allowing for more nuanced descriptions of emotional causes. Our rigorous annotation process involves five graduate students specializing in sentiment analysis. Two annotators independently label each instance, and their annotations are evaluated using BERTScore [69]. When the BERTScore exceeds 0.75, the remaining three students vote to determine the final annotation. Otherwise, annotators discuss and revise until a consensus is reached. This protocol, enhanced by automatic evaluation metrics, ensures higher annotation reliability compared to ECGF's approach, which lacks systematic consistency checks between annotators.

To address the MECEC task, we propose FAME-Net (Facial-Aware Multimodal Emotion Explanation Network), a novel architecture built upon LLaVA [41]. Drawing from research showing that facial expressions provide crucial cues for emotion cause identification [14, 26], FAME-Net incorporates both visual and facial emotion recognition to capture subtle emotional dynamics often missed by text-only analysis. FAME-Net features a specialized two-stage facial emotion recognition pipeline, first detecting and tracking faces within video scenes, then extracting facial emotion features through multi-scale feature networks and global depth convolution techniques. This comprehensive multimodal approach enables FAME-Net to identify both explicit and implicit emotional triggers, achieving strong performance across standard text generation metrics.

Our main contributions are as follows:

- We propose a new Emotion Cause Explanation dataset (ECEM) based on MELD. This dataset provides detailed natural language explanations for the emotion causes of the target utterances and is specifically designed for the training and evaluation of MECEC tasks.
- We propose FAME-Net, integrating visual modality and facial emotion recognition for comprehensive emotional cause analysis. Through a two-stage pipeline, it accurately extracts emotional features, improving multi-modal emotion cause explanation and generation performance.
- The experimental results on the ECGF and ECEM datasets demonstrate that our proposed model, FAME-Net, surpasses existing methods across the majority of evaluation metrics.

2 Related Works

2.1 Causal Emotion Entailment

Over the past decade, emotion cause analysis has gained attention for enhancing emotional intelligence, focusing on two key tasks: Emotion Cause Extraction (ECE) and Emotion Cause Pair Extraction (ECPE). Initially proposed by Lee et al. [33], ECE identifies text segments explaining emotion causes. Gui et al. [23] later refined ECE at the clause level using a Chinese dataset. Poria et al. [48] advanced this with dual-stream attention (TSAM) and commonsense integration. While existing tasks often use multiple-choice formats, generative causal reasoning remains underexplored. Xia and Ding [62] proposed ECPE to jointly extract emotions and causes, broadening applications. Wang et al. [57] introduced Multimodal ECPE (MECPE) by incorporating text, audio, and video. Recently, Wang et al. [58] proposed MECGC for fine-grained causal explanations and MECEC, a generative task ensuring conversational continuity in multimodal dialogues. These advancements highlight the field's evolving potential.

2.2 Large Language Models

The advent of Large Language Models (LLMs) has greatly impacted Natural Language Processing (NLP) and multimodal learning, revolutionizing artificial intelligence. LLMs like OpenAI's GPT [7] excel at a variety of language tasks, including translation, summarization, question answering, and content generation. These models are trained on vast data, allowing them to perform with impressive accuracy and engage in complex conversations and reasoning,

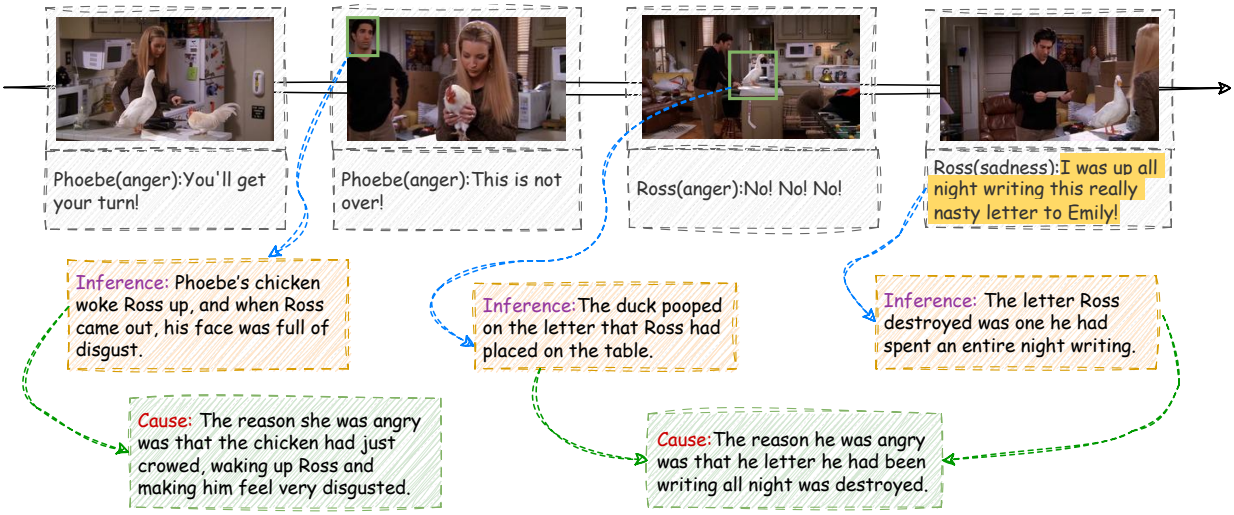


Figure 2: An example of the annotated conversation in our ECEM dataset. Inference represents the reasoning conducted during the annotation process, while cause, represents the emotional causes finally annotated in the dataset.

previously thought to be human-exclusive. This shift has transformed NLP applications, from research to commercial uses such as chatbots and personalized recommendations. Simultaneously, models like CLIP [50] and BLIP [34] have advanced the integration of visual and textual understanding. CLIP can process both images and text, enabling tasks like image search and caption generation. BLIP further refines this, advancing visual-textual applications in areas like visual storytelling. Multimodal learning now includes dynamic media like video and audio. Models such as LLaVA [42] and Video-ChatGPT [44] combine video and audio analysis with AI systems, enhancing interactions and improving applications like video summarization and real-time communication. This shift reflects the growing need for more sophisticated AI that mimics human communication across multiple senses. FAME-Net can take into account the emotional factors of the individuals in the video while processing visual information.

3 Dataset Construction

3.1 Annotation Program

We first provide a detailed explanation of the dataset annotation method. To ensure consistency in style and accuracy of the annotated data, we refer to the research findings [15, 39]. Specifically, we invite five graduate students specializing in sentiment analysis to participate in annotation work. We first train them on annotation methods, ensuring they correctly understand the task definition and proficiently master the use of annotation tools, thereby guaranteeing the quality of the dataset we construct. Then, two students independently annotated the entire dataset. We calculate the BERTScore [69] of their annotations. For cases with scores exceeding 0.75, the remaining three graduate students independently evaluate the annotations, with the final result determined by majority vote. When scores are below 0.75, the annotators discuss discrepancies to reach consensus. This ensures the rationality and consistency of data annotation. During the data annotation process, we observe issues such as missing textual data relative to the video content and incorrect use of punctuation. Therefore, in our experiments, we used a combination of WhisperX [4] and WhisperAT [21] to transcribe audio information from videos to correct the original data and improve data quality.

Data Pre-Processing. We preprocess the MELD dataset by merging preceding video clips into a continuous sequence to accurately capture emotional causes of each utterance without interference from future content. This method is beneficial for elucidating the emotional causes associated with the final discourse in the video segments. For example, in a dialogue consisting of two utterances, U_0 and U_1, the concatenated results would be (U_0) and (U_0, U_1). During the concatenation process, overlapping segments between the original video clips resulted in redundant content. Therefore, we design an algorithm to detect and remove duplicate frames to enhance the quality of our dataset.

Annotation Guidelines. We focus more on different emotions; therefore, this paper does not involve the neutral emotions. Before the formal annotation, we randomly assign 100 videos to annotators for training, and formal annotation only begins after the training results receive unanimous approval from evaluators. For each non-neutral utterance, annotators are required to find corresponding evidence from the three modalities of the dialogue and briefly describe their reasoning in one to three sentences. We specifically request that annotators pay close attention to the video and audio modalities to ensure the accuracy and completeness of the reasons provided. Annotated examples are shown in the Figure 2.

3.2 Data Statistics and Analysis

Comparison with Existing Datasets and Data Analysis. As shown in Table 1, by comparing the ECEM dataset with mainstream datasets

Table 1: Comparison of Cause Detection Datasets, including traditional Emotion Cause Analysis (ECA), Emotion Cause Detection (ECD), Sarcasm Detection (SD), and Humor Detection (HD), etc. The units of the instances include sentence (s), document (d), post (p), utterance (u) and video (v). T, A, V stands for text, audio and video.

Dataset	Modality			Task	Cause type	Scene	Ins
	T	A	V				
ECE Corpus [24]	✓	✗	✗	ECE	Extractive	News	2,105 d
NTCIR-13-ECA [19]	✓	✗	✗	ECA	Extractive	Fiction	2,403 d
Weibo-Emotion [11]	✓	✗	✗	ECD	Extractive	Blog	7,000 p
REMAN [29]	✓	✗	✗	ECD	Extractive	Fiction	1,720 d
GoodNewsEveryone [6]	✓	✗	✗	ECD	Extractive	News	5,000 s
MHD [47]	✓	✗	✓	HD	Extractive	Sitcom	13,633 u
UR-FUNNY [25]	✓	✓	✓	HD	Extractive	Speech	16,514 u
MUStARD [8]	✓	✓	✓	SD	Extractive	Sitcom	6,365 v
WITS [31]	✓	✓	✓	SD	Generative	Sitcom	2,240 u
SMILE [27]	✓	✓	✓	HD	Generative	Sitcom&Talks	887 v
ExFunTube [30]	✓	✓	✓	HD	Generative	Youtube	10,136 v
RECCON-IE [48]	✓	✗	✗	ECE	Extractive	Conv	665 u
RECCON-DD [48]	✓	✗	✗	ECE	Extractive	Conv	11,104 u
ConvECPE [35]	✓	✗	✗	ECPE	Extractive	Conv	7,433 u
ECF [57]	✓	✓	✓	MECPE	Extractive	Conv	13,619 u
ECGF [58]	✓	✓	✓	MECGC	Generative	Conv	13,619 u
ECEM (ours)	✓	✓	✓	MECEC	Generative	Conv	7,273 u

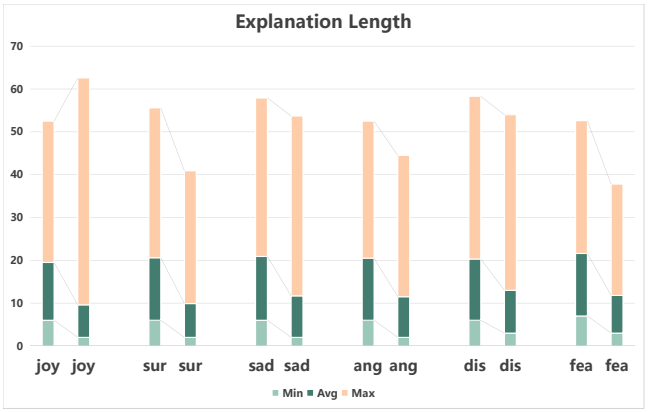


Figure 3: Statistical analysis of the reason length for emotion causes is conducted on the datasets ECEM and ECGF. The left side represents ECEM, while the right side represents ECGF.

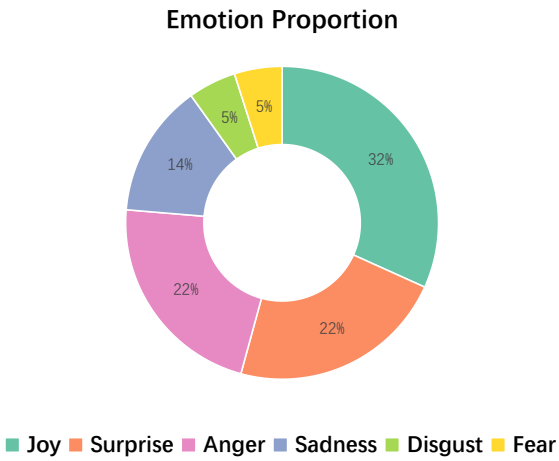


Figure 4: Proportion of six classified emotions.

in the field of emotion cause detection, we find that existing datasets (such as those for humor and sarcasm detection) are often limited to single emotions or fragmented explanations. Although the ECGF [58] dataset supports multi-emotion explanations, its annotation method based on video segmentation results in excessively short explanations—averaging 8.64 words. In contrast, the ECEM dataset not only covers six emotions but also achieves more comprehensive emotion attribution through coherent explanations averaging 14.4 words. The comparison of explanation lengths between ECGF and ECEM is illustrated in Figure 3. To ensure data quality, we remove neutral emotion samples during the construction process and conduct statistical analysis on the proportions of the six emotions, with the results shown in Figure 4. Additionally, we perform a statistical analysis of the annotation text lengths, revealing that explanations for different emotions exhibit high consistency in terms of maximum, minimum, and average values. This standardized pattern not only validates the rigor of the annotation process but also confirms the dataset's reliability in cross-emotion explanation capabilities. This dual advantage—surpassing the structural limitations of existing datasets while incorporating a rigorous data validation mechanism—makes ECEM the most interpretable benchmark for emotion cause analysis to date.

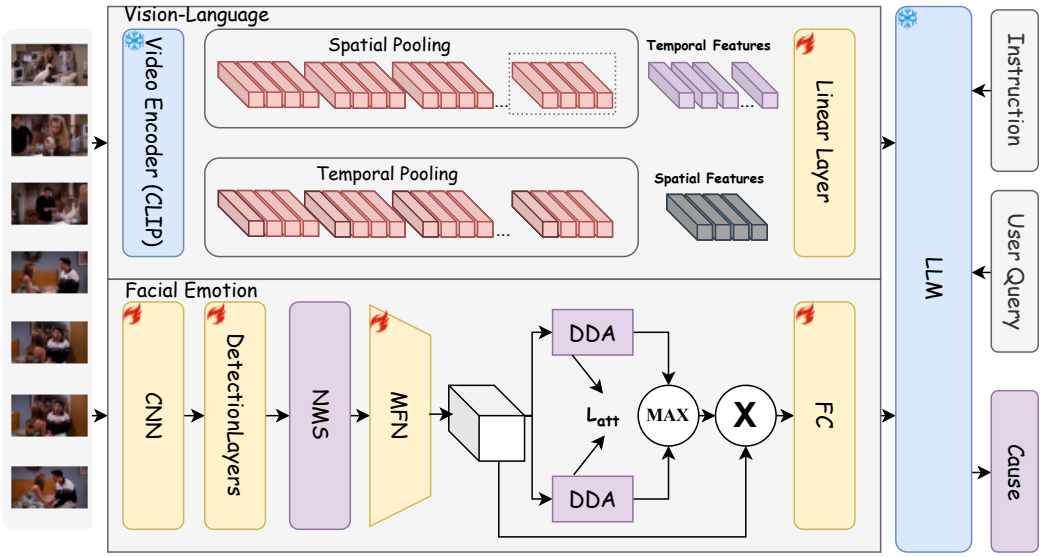


Figure 5: Overview of our proposed FAME-Net. The MFN network includes multiple convolutional layers, depthwise convolutional layers, residual connections, and a final linear layer. DDA includes convolution, attention mechanisms, and linear transformations. X represents multiplication.

4 Method

4.1 Overview

We present the Facial-Aware Multimodal Emotion Explanation Network (FAME-Net), a multimodal large language model (LLM) built on the LLaVA framework [42]. LLaVA is selected for its optimal balance between computational efficiency and model performance, as well as its adaptable interfaces for video-related tasks, which allow researchers to prioritize refining temporal modeling and advancing complex reasoning mechanisms. As illustrated in Figure 5, FAME-Net integrates visual, textual, and facial emotion modalities. Within its visual-language module, spatiotemporal feature fusion is employed to concurrently capture temporal dynamics and spatial structural patterns in video data, enhancing fine-grained content understanding. Facial emotion recognition complements text-based analysis limitations [14], quantifies emotional contagion to uncover latent affective drivers [26], and detects discrepancies between facial expressions and verbal cues to identify concealed emotional states [9]. Our work focuses on modeling the temporal evolution and contextual dependencies of facial expressions in video streams, proposing a novel architecture tailored for dynamic emotion analysis.

During its development, FAME-Net integrates advanced mechanisms to align visual signals with language processing, enhancing multimodal analysis. The model features a CLIP-based video encoder [50] tailored to capture the information from spatial and temporal dimensions of video data, alongside a facial emotion recognition module that accurately identifies and interprets subtle emotional nuances. This combination enables FAME-Net to reveal deeper emotional insights beyond conventional text analysis.

4.2 Vision-Language Module

Our architecture, as shown in the "Vision-Language" section of Figure 5, employs the CLIP ViT-L/14@336 as the visual encoder, adapting it from image processing to video processing to meet the requirements of FAME-Net. This adaptation is crucial for the model to effectively capture spatio-temporal representations in videos. For a video, $V = \{I_1, ..., I_k, ..., I_T\}$, where $I_k \in \mathbb{R}^{H\times W\times C}$ and T denotes the number of frames. The encoder independently processes each frame, generating the frame-level embedding $x_k = CLIP(I_k), x_k \in \mathbb{R}^{h\times w\times D}$, where $h = H/p$, $w = W/p$ and p represents the patch size, and $h \times w = N$. The video embedding can be obtained, $X = [x_1, .., x_k, ..., x_T]$.

To construct a comprehensive video-level representation, we apply average pooling along the temporal dimension of frame-level embeddings, resulting in a temporal representation $R_t \in \mathbb{R}^{T\times D}$:

$$R_t = \frac{1}{N}\sum_{i=1}^{N} X \tag{1}$$

This temporal pooling technique effectively integrates information across frames. Similarly, we apply average pooling along the spatial dimension to obtain the spatial representation $R_s \in \mathbb{R}^{N\times D}$:

$$R_s = \frac{1}{T}\sum_{i=1}^{T} X \tag{2}$$

The final video-level features v_j are a concatenation of these temporal and spatial features.

$$R = [R_t \oplus R_s] \in \mathbb{R}^{(T+N)\times D} \tag{3}$$

where $\oplus$ is the concatenation operation.

A simple trainable linear layer g, projects these video-level features into the language decoder's embedding space, transforming them into corresponding language embedding tokens Q_v.

$$Q_v = g(R) \in \mathbb{R}^{(T+N)\times K} \tag{4}$$

where K represents the dimension of the text query.

Specific inquiries regarding emotional causes for explanations: User Query is denoted as $Q_t \in \mathbb{R}^{L\times K}$ where L is the length of the query. These queries are tokenized to be dimensionally compatible with the video embeddings. Finally, Q_v and Q_t are input to the language decoder together.

4.3 Facial Emotion Module

As shown in the "Facial Emotion" section of Figure 5, we further design a two-stage system, where the first stage performs face detection to segment faces from the video, and the second stage conducts facial emotion recognition. This two-stage detection system reduces the impact of the environment on facial emotion recognition and improves the accuracy of detection.

In the first stage, in order to efficiently extract features and adapt to various scale changes in face, we use CNN [32] to extract feature maps F from the input video frame I, and then, a set of multi-scale anchors is generated on the feature map, and these anchors a are classified using Detection Layers composed of VGG16 [53] and CNNs to determine whether they contain a face. Simultaneously, bounding box regression is applied to adjust the position and size of the candidate detection boxes. The process of classification and regression can be represented as:

$$\mathbb{1}_{face},\ \Delta = DetectionLayers(F,\ a) \tag{5}$$

where $\mathbb{1}_{face}$ represents whether it contains a face, and Δ represents the bounding box score. Using the regression output Δ to adjust the position and size of each anchor, a, the detection box, B, is generated. To reduce the probability of detecting the same face multiple times, we use Non-Maximum Suppression (NMS) [17] with an Intersection over Union (IoU) threshold to obtain the final detected faces.

$$B_{final} = \mathrm{NMS}(B,\ \mathbb{1}_{face},\ IoU_{threshold}) \tag{6}$$

In the second phase, to reduce the loss of spatial information and retain more contextual details, we first use a MFN [66] to extract feature maps F_{MFN} from the input face B_{final}:

$$F_{\mathrm{MFN}} = \mathrm{MFN}(B_{final}) \tag{7}$$

The features of facial emotions are not limited to a single direction. By considering both directions simultaneously, the model can more comprehensively understand the complexity of facial expressions. Therefore, we use DDA [68] to generate attention maps in both the vertical direction (Y) and the horizontal direction (X).

$$\begin{aligned} F^X_{\mathrm{DDA}} &= \mathrm{DDA}_X(F_{\mathrm{MFN}}) \\ F^Y_{\mathrm{DDA}} &= \mathrm{DDA}_Y(F_{\mathrm{MFN}}) \end{aligned} \tag{8}$$

We integrate attention mechanisms from two directions to enhance recognition capability and accuracy. The sigmoid function, σ, is used to suppress noise while strengthening the critical areas of the features. This result is multiplied by the max-pooling output to further enhance important features and suppress less significant information, yielding the final attention feature map F_{att}.

$$F_{\mathrm{att}} = pooling(F^X_{\mathrm{DDA}}, F^Y_{\mathrm{DDA}}) \cdot \sigma(F^X_{\mathrm{DDA}} F^Y_{\mathrm{DDA}}) \tag{9}$$

Finally, the attention feature map is passed through a fully connected layer to predict the emotion category of the face, E_{cls}.

$$E_{cls} = FC(F_{\mathrm{att}}) \tag{10}$$

4.4 Video Instruction Tuning

During the fine-tuning phase, we use predefined prompts, $\mathtt{Prompt}_{\mathrm{Video}}$.

```
USER: <Instruction> <User Query><Vid-tokens><Emotion>
Assistant: Using the notations, we can represent it as:
USER: <Q_t> <Q_v> <E_cls>
Assistant:
```

where <Instruction> defines the task, which requires explaining the emotional reasons by understanding the video. <User Query> includes the dialogue history in multimodal conversations, as well as the utterance requiring emotional cause explanation and the emotion it conveys. The predicted answer corresponds to the final cause C. Throughout the training process, the weights of the video encoder and LLM remain freezed, and the model maximizes the likelihood of tokens representing the answer by adjusting the linear layer. <Q_t> represent <Instruction> and <User Query>; <Q_v> represent <Vid-tokens>.

5 Experiments

5.1 Implementation Details

Our FAME-Net model is built on the LLaVA-7b [41] foundation. To enhance training efficiency, we preprocess the faces in the video before training. The training process is conducted on two NVIDIA A6000 48G GPUs.

5.2 Evaluation Metrics

To comprehensively evaluate the fluency, diversity, and accuracy of generated causeing, we adopt a series of widely recognized automatic evaluation metrics for text generation, including BLEU [46], METEOR [5], ROUGE [38], and CIDEr [56], which focus on lexical and syntactic matching, as well as BERTScore [69] and BLEURT [52], which assess whether the generated text is semantically consistent with the reference text.

5.3 Baselines

We compare it with several open-source models of similar scale. Ultimately, we select NExT-GPT [61], ChatUniVi [28], Video-LLaMA[67], Qwen2-VL [59], LLaVA-NeXT-Video [40], and Video-LLaMA2 [12] as baselines. Also, we choose models like Qwen [3], Qwen1.5 [3], Qwen2 [65], Baichuan2 [64], LLaMA2 [55], LLaMA3 [22], and ChatGLM3 [2], ChatGLM4 [20], which excel in single text modality, for experiments where only conversation history is input. In addition, we select four pre-trained small models as baselines. The four small models are VALOR [43], COSA [10], ObG [58], and IGV [37]. We train and test these models on the ECEM dataset. For the ECGC dataset, we use the experimental results from the original paper as baselines.

Table 2: Comparison of Automatic Evaluation Metrics Experiments On ECEM.

Class	Method	BLEU4	METEOR	ROUGE	BERTScore	BLEURT	CIDEr
Pre-trained Small Models	IGV	0.0082	0.1201	0.2190	0.7920	-1.5200	0.5146
	ObG	0.0134	0.1791	027880	0.8131	-1.1456	0.5874
	COSA	0.0064	0.1261	0.2278	0.7674	-1.3576	0.5710
	VALOR	0.0085	0.1311	0.2581	0.7964	-1.2127	0.5264
LLMs	Qwen	0.0105	0.1432	0.1222	0.8498	-1.0425	0.5342
	Qwen1.5	0.1124	0.2775	0.3157	0.7693	-0.9878	0.4759
	Qwen2	**0.1235**	0.3221	0.4045	0.8685	-0.6177	0.6638
	Baichuan2	0.0245	0.1354	0.1577	0.8246	-1.1124	0.5344
	LLaMA2	0.0106	0.1519	0.2443	0.8507	-1.1445	0.5543
	LLaMA3	0.0117	0.1499	0.2758	0.8617	-0.9545	0.5698
	ChatGLM3	0.0007	0.1857	0.1258	0.8587	-0.7099	0.6049
	ChatGLM4	0.0010	0.1694	0.1294	0.8629	-0.6921	0.6196
MLLMs	ChatUniVi	0.0136	0.1331	0.1501	0.8245	-1.1092	0.4752
	Qwen2-VL	0.0861	0.3214	0.3634	0.8721	-0.5463	0.6557
	NExT-GPT	0.0945	0.1250	0.3401	0.7599	-1.2821	0.6397
	Video-LLaMA	0.0871	0.3324	0.3724	0.8981	-0.5463	0.6677
	Video-LLaMA2	0.0012	0.1785	0.1300	0.8361	-0.8779	0.6268
	LLaVA-NeXT-Video	0.0654	0.3235	0.3966	0.8974	-0.6634	0.6788
	FAME-Net (ours)	0.0870	**0.3423**	**0.4051**	**0.9053**	**-0.4789**	**0.6836**

Table 3: Comparison of Automatic Evaluation Metric Experiments On ECGC.

Modality	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	BERTScore
Text	GPT-3.5	0.2413	0.1582	0.1157	0.0886	0.1722	0.2258	0.8518	0.6801
	Gemini-Pro	0.2624	0.1726	0.1272	0.0984	0.1767	0.2497	0.8926	0.6936
	T5s	0.4815	0.4435	0.3718	0.3403	0.2979	0.4608	2.8770	0.7619
	Flan-T5	0.4897	0.4212	0.3804	0.3499	0.3057	0.4751	3.0279	0.7698
MM	Gemini-Pro-Vision	0.2780	0.1826	0.1371	0.1085	0.1798	0.2453	0.8960	0.6960
	ObG	**0.5011**	**0.4341**	**0.3939**	**0.3641**	0.3008	0.4712	3.0079	0.7672
	FlanObG	0.4967	0.4313	0.3924	0.3631	0.3042	0.4781	3.0808	0.7711
	FAME-Net (ours)	0.4578	0.4214	0.3921	0.3587	**0.3150**	**0.4786**	**3.1347**	**0.8454**

5.4 Main Results and Analysis

5.4.1 Automatic Evaluation. Table 2 presents the performance of the baseline model and FAME-Net on the ECEM dataset across six evaluation metrics. From the table, it is evident that the FAME-Net model excels in multiple key metrics, particularly achieving the best performance in METEOR, ROUGE, BERTScore, BLEURT, and CIDEr. These metrics outperform other comparative models, including pre-trained small models, Large Language Models (LLMs), and Multimodal Large Language Models (MLLMs). Although FAME-Net does not achieve the highest score in the BLEU4 metric, this does not detract from its overall superior performance. BLEU4, as an evaluation metric based on n-gram overlap, can reflect the local accuracy of generated text in some cases, but its ability to capture semantic understanding and contextual coherence is limited. In contrast, other metrics place more emphasis on semantic similarity and contextual consistency. Therefore, the excellent performance of FAME-Net in these metrics fully demonstrates its strong capability in the task of MECEC, particularly highlighting its significant advantages in semantic understanding and generation quality.

Table 3 shows the scores of the FAME-Net model on the ECGC dataset, where GPT-3.5 [45], Gemini-Pro [54], T5 [51], and Flan-T5 [51] are pure text models, and Gemini-Pro-Vision [54], ObG [58], and FlanObG [58] are multimodal models. The experimental results show that FAME-Net performs weakly on the BLEU metric but leads on metrics such as METEOR, ROUGE-L, CIDEr, and BERTScore. This indicates that our model has superior capabilities in semantic understanding, fluency, and creativity. Since the BLEU metric mainly focuses on n-gram matching, it may not fully reflect the model's advantages in language generation, while other metrics emphasize the semantic consistency and contextual understanding of the text. Therefore, despite the lower BLEU score, the model's outstanding performance in generation quality and expressive ability still indicates its higher potential in practical applications.

Overall, the FAME-Net model performs excellently on ECGF and ECEM datasets, showing leading advantages in multiple metrics and achieving a good balance between innovation and accuracy. In addition, both datasets demonstrate that multimodal models, including FAME-Net, outperform pure text models and other smaller models in overall performance. This indicates that integrating information from multiple models enhances the model's ability to solve MECEC task.

Table 4: Ablation Study Results, where V represents the visual modality, T represents the text modality, and E represents the facial emotion recognition module.

Method	BLEU4	METEOR	ROUGE	BERTScore	BLEURT	CIDEr
V	0.0375	0.2217	0.2455	0.8850	-0.9061	0.5572
V+E	0.0695	0.3058	0.3427	0.8960	-0.7715	0.6259
V+T	0.0700	0.3233	0.3771	0.8977	-0.7215	0.6626
V+T+E	**0.0870**	**0.3423**	**0.4051**	**0.9053**	**-0.4789**	**0.6836**

Table 5: Comparison between original data and transcription results.

Method	BLEU4	METEOR	ROUGE	BERTScore	BLEURT	CIDEr
Original Text	0.0495	**0.3425**	0.4040	0.8978	-0.5031	0.6755
Improved Text	**0.0870**	0.3423	**0.4051**	**0.9053**	**-0.4789**	**0.6836**

5.4.2 Human Evaluation and Analysis. In addition to automatic evaluation, we randomly select 100 test samples from the four best-performing models for manual evaluation. We choose three graduate students specializing in sentiment analysis as evaluators and ask them to score the generated reasons based on three criteria: coherence, fluency, and relevance, with a rating scale from 1 to 5. Coherence emphasizes the internal consistency of generated text in terms of logic, semantics, and structure. Fluency evaluates the grammatical correctness and naturalness of language use in the sentences. Relevance determines whether the generated sentences contain the exact reasons for expressing emotions. As shown in Table 6, FAME-Net achieves the best performance across all three metrics. This result demonstrates that FAME-Net not only excels in the logical consistency and linguistic naturalness of generated text but also more accurately captures the reasons related to emotional expression, showcasing its strong capability in the MECEC task.

5.5 Ablation Experiments

The ablation experiment results, as shown in Table 5, confirm that multimodal information fusion enhances dialogue generation performance. The joint visual-text modeling improves the average performance across all metrics by 7.1% compared to the pure visual model, demonstrating the effectiveness of cross-modal semantic alignment. The introduction of the facial emotion module increases METEOR and BLEURT by 2.5% and 7.3%, respectively, indicating that emotional features can enhance the empathetic ability of expressions. The final V+T+E model achieves optimal performance across all evaluation metrics, validating the synergistic effect of visual content understanding, textual semantic reasoning, and emotional feature capturing. This proves that complementary multimodal information can more comprehensively support dialogue generation tasks.

5.6 Original Data and Transcription

To verify our hypothesis that audio transcription data is of higher quality compared to the original text data, we conduct a detailed and thorough comparative experiment, with the results comprehensively displayed in Table 5. The BLEU4 score is higher for the transcription, indicating that the transcribed text aligns more closely and accurately with the reference standard in terms of lexical matching, thereby demonstrating superior accuracy and consistency. Although the METEOR score remains nearly identical between the two datasets, the transcription results still slightly outperform the original text across all other evaluation metrics, suggesting an overall enhancement and refinement in data quality following transcription.

Table 6: Human evaluation results, Coh, Flu, Rel represents Coherence, Fluency, Relevance.

Method	Coh	Flu	Rel
LLaVA-NeXT-Video	3.91	4.14	3.85
Video-LLaMA	4.05	4.24	3.99
Video-LLaMA2	3.97	4.17	3.89
FAME-Net (ours)	**4.15**	**4.31**	**4.01**

6 Conclusion

Our paper constructs a novel dataset, ECEM, for training and evaluating the MECEC task. To address the MECEC task, we propose a specialized model, FAME-Net, based on the LLaVA framework. FAME-Net excels in understanding video content and provides natural language explanations for the emotions of characters depicted in the videos. Extensive experimental results demonstrate that our model performs exceptionally well in the MECEC task. Specifically, FAME-Net, by integrating multimodal information, captures emotional changes in characters more accurately and generates coherent and emotionally rich explanations. Additionally, the construction of the ECEM dataset considers diversity and complexity, covering various emotion types and scenarios, ensuring the model's generalization ability in practical applications.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (62272092,62172086) and the Fundamental Research Funds for the Central Universities of China (No.N2116008).

References

[1] Christopher R. Agnew and Susan C. South. 2014. *Interpersonal Relationships and Health: Social and Clinical Psychological Mechanisms.* Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780199936632.003.0001

[2] Zhipu AI and Tsinghua University KEG. 2023. ChatGLM3: A large-scale Chinese dialogue language model. https://github.com/THUDM/ChatGLM3.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[4] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747* (2023).

[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* 65–72.

[6] Laura Bostan, Evgeny Kim, and Roman Klinger. 2019. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *arXiv preprint arXiv:1912.03184* (2019).

[7] 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 (2020), 1877–1901.

[8] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).

[9] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[10] Sihan Chen, Xingjian He, Handong Li, Xiaojie Jin, Jiashi Feng, and Jing Liu. 2023. Cosa: Concatenated sample pretrained vision-language foundation model. *arXiv preprint arXiv:2306.09085* (2023).

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

[12] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476* (2024).

[13] Hilda Costa, Pilar Ripoll, Miguel Sánchez, and Carla Carvalho. 2013. Emotional intelligence and self-efficacy: Effects on psychological well-being in college students. *The Spanish Journal of Psychology* 16 (2013), E50.

[14] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[15] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1473–1482.

[16] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. Empathyear: An open-source avatar multimodal empathetic chatbot. *arXiv preprint arXiv:2406.15177* (2024).

[17] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32, 9 (2009), 1627–1645.

[18] Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021.* 807–819.

[19] Qinghong Gao, Jiannan Hu, Ruifeng Xu, Lin Gui, Yulan He, Kam-Fai Wong, and Qin Lu. 2017. Overview of NTCIR-13 ECA Task.. In *NTCIR.*

[20] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

[21] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. *arXiv preprint arXiv:2307.03183* (2023).

[22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[23] Lin Gui, Ruifeng Xu, Qin Lu, Dongyin Wu, and Yu Zhou. 2016. Emotion cause extraction, a challenging task with corpus construction. In *Social Media Processing: 5th National Conference, SMP 2016, Nanchang, China, October 29–30, 2016, Proceedings.* Springer, 98–109.

[24] Lin Gui, Ruifeng Xu, Dongyin Wu, Qin Lu, and Yu Zhou. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond.* World Scientific, 145–160.

[25] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618* (2019).

[26] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science* 2, 3 (1993), 96–100.

[27] Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2023. SMILE: Multimodal Dataset for Understanding Laughter in Video with Language Models. *arXiv preprint arXiv:2312.09818* (2023).

[28] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv preprint arXiv:2311.08046* (2023).

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

[30] Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can Language Models Laugh at YouTube Short-form Videos? *arXiv preprint arXiv:2310.14159* (2023).

[31] Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419* (2022).

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[33] Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text.* 45–53.

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning.* PMLR, 12888–12900.

[35] Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2022. ECPEC: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing* 14, 3 (2022), 1754–1765.

[36] Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 2041–2045.

[37] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2928–2937.

[38] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out.* 74–81.

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.* Springer, 740–755.

[40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

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

[43] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

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

[45] OpenAI. 2023. ChatGPT (GPT-3.5). https://chat.openai.com Accessed: 2025-01-24.

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 311–318.

[47] Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 576–585.

[48] 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 (2021), 1317–1332.

[49] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

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

[51] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[52] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[53] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[54] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[57] Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2022. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing* 14, 3 (2022), 1832–1844.

[58] Fanfan Wang, Heqing Ma, Xiangqing Shen, Jianfei Yu, and Rui Xia. 2024. Observe before Generate: Emotion-Cause aware Video Caption for Multimodal Emotion Cause Generation in Conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5820–5828.

[59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[60] Mi-Young Woo, Hyun-Jii Bae, and Sung-Jung Hong. 2020. The convergence effect of emotional intelligence and interpersonal relationship competence on communication skill of nursing students. *Journal of the Korea Convergence Society* 11, 12 (2020), 401–408.

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

[62] Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267* (2019).

[63] Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A RNN-transformer hierarchical network for emotion cause extraction. *arXiv preprint arXiv:1906.01236* (2019).

[64] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305* (2023).

[65] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[66] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

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

[68] Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. 2023. A dual-direction attention mixed feature network for facial expression recognition. *Electronics* 12, 17 (2023), 3595.

[69] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[70] Xiaopeng Zheng, Zhiyue Liu, Zizhen Zhang, Zhaoyang Wang, and Jiahai Wang. 2022. Ueca-prompt: Universal prompt for emotion cause analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*. 7031–7041.