Locate and Explain: Joint Multimodal Emotion Cause Extraction and Summarization in Conversation

Jikun Wan, Chen Gong*, Guohong Fu
Institute of Artificial Intelligence, School of Computer Science and Technology,
Soochow University, Suzhou, China
{wanjikun1135}@gmail.com;
{gongchen18, ghfu}@suda.edu.cn

Abstract

Multimodal emotion cause analysis in conversation aims to identify the causes of emotions by leveraging multimodal information. Existing studies mainly formulate this problem as either utterance-level emotion cause extraction, which provides clear cause localization but limited explanation, or multimodal emotion cause generation, which offers fine-grained explanations but lacks explicit traceability to source utterances. Moreover, existing datasets rely heavily on human judgment and lack well-defined structured theoretical guidance, leading to subjective and inconsistent annotations. To address these issues, we introduce joint Multimodal Emotion Cause Extraction and Summarization in conversation (MECES), a new task that simultaneously extracts emotion cause utterances and generates cause summaries, enabling both precise localization and interpretable explanations of emotion cause. We further construct a MECES dataset guided by the Activating events–Beliefs–Consequences theory from psychology. This dataset consists of 5,787 emotion utterances annotated with causes, comprising 12,231 emotion-cause pairs and 6,040 cause summaries. We also propose an effective end-to-end joint learning approach for MECES task, establishing strong benchmark results for this newly introduced task and dataset.

1 Introduction

Multimodal Emotion Cause Analysis in Conversation (MECAC) aims to identify the causes of emotions in conversations by leveraging textual, acoustic, and visual cues. Understanding these causes is essential for building dialogue systems that respond empathetically and behave more human-like (Fei et al., 2024; Wang et al., 2024).

Existing MECAC research mainly follows two task formulations. The first formulation, utterance-level emotion cause extraction (Wang et al., 2023; Liang et al., 2025; Wu et al., 2025), identifies which utterances trigger the emotion in a target utterance, providing clear cause localization. For example, in Figure 1, Guang Shi's sad emotion in Utterance 4 (U4) is caused by receiving a practice workbook as a gift, mainly conveyed visually in U4, so U4 is marked as the cause. However, extraction methods do not explain what in the utterance actually triggers the emotion. To address this, the second formulation, multimodal emotion cause generation (Wang et al., 2024) produces summaries that describe causes in more detail. For instance, the cause of the sad emotion in U4 is summarized as "Guang Shi realizes that the gift from his teacher is a Go book, implying an unexpected expectation to continue studying." While this cause generation approach provides fine-grained explanations, it lacks explicit source attribution, making the causes difficult to verify.

In addition to task-level limitations, existing MECAC datasets face several challenges. The annotation process relies heavily on human judgment and lacks well-defined structured theoretical guidelines, leading to subjective and inconsistent annotations. Moreover, most datasets constrain emotion causes to single factors, such as objective events or subjective beliefs (Wang et al., 2023; Liang et al., 2025), overlooking the complex interactions that often contribute to emotion generation.

In this work, we introduce a new task, joint **M**ultimodal **E**motion **C**ause **E**xtraction and **S**ummarization in conversation (MECES), which aims to simultaneously extract emotion cause utterances and generate abstractive cause summaries. By doing so, MECES not only explains what the emotion cause is, but also clarifies where the supporting evidence lies, as illustrated in Figure 1.

To support this task, we construct a joint multimodal emotion cause extraction and summarization dataset, namely MECESD. The annotation process is guided by the Activating

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43472–43489

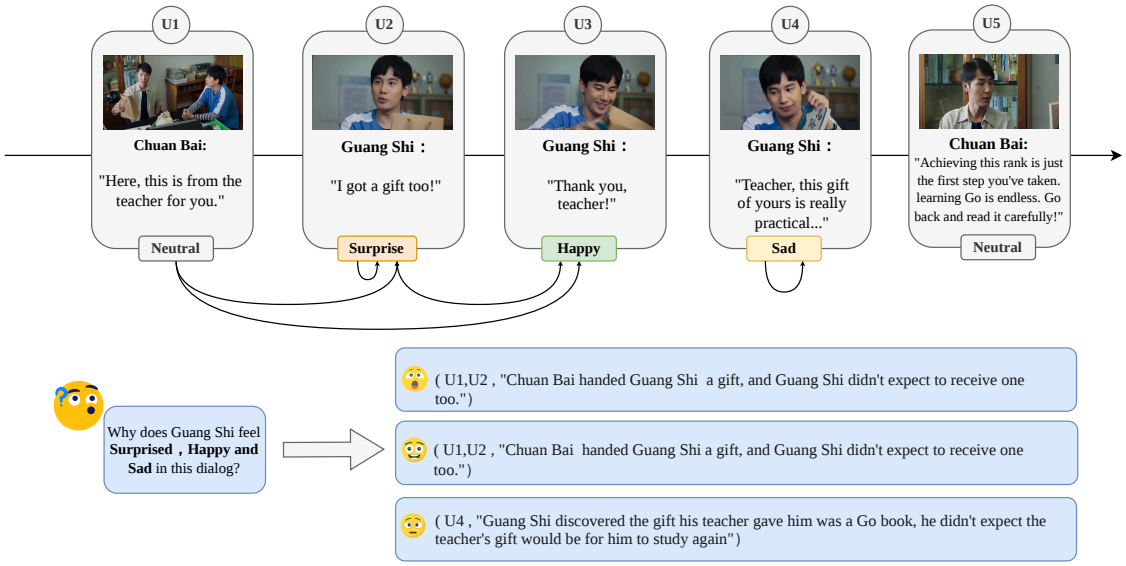

Figure 1: An example of the MECES task. The arcs in the upper part point from the emotion cause utterances to the resulting emotion.

events–Beliefs–Consequences (ABC) theory from psychology (Ellis, 1957), which helps ensure both the comprehensiveness of emotion attribution and annotation consistency. MECESD consists of 5,787 emotion utterances annotated with causes, comprising 12,231 emotion-cause pairs and 6,040 cause summaries, where a single utterance may be associated with multiple valid emotion cause summaries.

Finally, we further propose a simple yet effective end-to-end joint learning approach for the MECES task to better integrate extraction and generation task. Extensive experiments on MECESD and other public datasets demonstrate the effectiveness of our approach, and establish strong benchmark results for this newly introduced task and dataset.

We will release the datasets and code at https://github.com/wwwper/MECES to facilitate future research.

2 Related Work

Existing MECAC research primarily centers on the Multimodal Emotion-Cause Pair Extraction (MECPE) task for utterance-level cause extraction (Wang et al., 2023; Liang et al., 2025). However, this utterance-level granularity is often considered too coarse. To address this limitation and capture more fine-grained details, the Multimodal Emotion Cause Generation in Conversation (MECGC) (Wang et al., 2024) task was introduced to directly produce cause summaries.

Although the MECGC task offers granular causal explanations through abstractive summarization, it suffers from a limitation: generated causes cannot be accurately traced back to specific source utterances. Consequently, we introduce a new task: MECES. This task aims to provide a comprehensive "localization and explanation" framework for emotion cause analysis in complex multimodal dialogues.

MECAC Datasets. Based on distinct MECAC task paradigms, existing datasets can be broadly categorized into two main groups. In the domain of cause extraction, Wang et al. (2023) first released the ECF dataset to establish a benchmark for the MECPE task. Subsequently, MECAD (Liang et al., 2025) and MEC^4 (Wu et al., 2025) were constructed to address issues related to data scarcity and cross-cultural challenges, respectively. Conversely, for the generative MECGC task, the field currently relies primarily on the ECGF (Wang et al., 2024) dataset to facilitate the research and evaluation of generative methods. However, these datasets suffer from limitations such as high subjectivity during the annotation process and incomplete attribution perspectives. Notably, we did not perform further annotation on the M^{3}HG dataset (Liang et al., 2025) due to two key differences: 1) Finer Granularity: We merged utterances based on semantic integrity rather than simple concatenation of speaker and emotion. This preserves finer details for identifying emotional causes. 2) Theoretical Framework: Our annotations are grounded in

the ABC theory, which differs from the framework used in M^{3}HG.

To address these limitations and support the MECES task, we propose MECESD, the first dataset jointly annotated for both extraction and summarization. By introducing the ABC theory for annotation, we effectively enhance annotation accuracy and ensure the comprehensiveness of attribution perspectives.

MECAC Methods. For the MECPE task, Li et al. (2024) proposed the HiLo model, which leverages emotion label constraints as cues to reinforce causal reasoning. Hu et al. (2024) formalized the task as a masked prediction problem, integrating multimodal semantic information via prompting. Liang et al. (2025) designed a multimodal heterogeneous graph approach to fuse multi-scale semantic information at both inter-utterance and intra-utterance levels. Wu et al. (2025) proposed the M^{3}F framework, which aggregates global temporal features of non-linguistic modalities by introducing long-term memory banks into a Q-Former, and leverages LLMs to achieve multimodal emotion cause pair extraction. For the MECGC task, ObG (Wang et al., 2024) stands as a representative work, generating emotion-cause-aware video descriptions to assist in the final cause generation.

While these methods excel in their respective domains, they generally address extraction or generation in isolation, thereby overlooking the intrinsic correlations between these two sub-tasks in MECES. To address this gap, we propose an end-to-end joint learning framework to perform both tasks simultaneously, aiming to facilitate mutual enhancement and provide reliable benchmark results for MECES.

3 Dataset Construction

In this section, we describe the annotation methodology and process for constructing the MECES dataset and present an in-depth data analysis.

3.1 Data Selection and Preprocessing

We select the publicly available M^{3}ED (Zhao et al., 2022) dataset as the initial data source. This dataset consists of dialogue clips collected from various Chinese television series, covering genres such as family, workplace, and urban life. Such diversity ensures broad coverage of real-world conversational scenarios. The public evaluation version of the dataset contains 811 dialogues.

To improve data quality and suitability for our task, we perform the following preprocessing steps. First, we merge fragmented utterances based on semantic integrity to address raw splits caused by pauses, ensuring that each utterance retains complete semantic and emotional cues. Second, we correct textual errors and align multimodal timestamps. Finally, we remove dialogues that lack sufficient contextual information for identifying emotion causes. After preprocessing, we obtain 781 high-quality multimodal dialogues comprising 10,136 utterances for annotation.

3.2 Annotation Guideline

In our annotation task, we aim to identify which utterances trigger the emotion expressed in a target utterance and to summarize the corresponding emotion cause. To better understand emotion causes and improve annotation quality and consistency, we take the Activating events-Beliefs-Consequences (ABC) (Ellis, 1957) theory from psychology as our theoretical guidance. As illustrated in Figure 2, the ABC theory posits that emotions arise from the interaction between both activating events and beliefs, rather than from a single factor, as is commonly assumed in previous datasets (Wang et al., 2023; Liang et al., 2025). For example, in Figure 2, Shanshan Zhu's sadness is triggered not solely by the activating event of being assigned a task in U2, but more critically by her belief that she lacks knowledge of renovation and therefore does not want to take on the task in U3. Had she instead believed that she was competent for the job, the same event might have elicited happiness rather than sadness. This example highlights that understanding emotion causes requires jointly considering both activating events and beliefs. Accordingly, the cause of Shanshan Zhu's sadness is attributed to both U2 and U3, and can be summarized as: "Sijin Fang asked Shanshan Zhu to accompany the renovation company to supervise the work, but Shanshan Zhu did not want to take on the task because she felt she lacked knowledge of renovation."

Guided by the ABC theory, we develop a detailed annotation guideline. This structured theoretical foundation improves annotation accuracy and consistency, while the dual-factor perspective ensures more comprehensive emotion cause attribution.

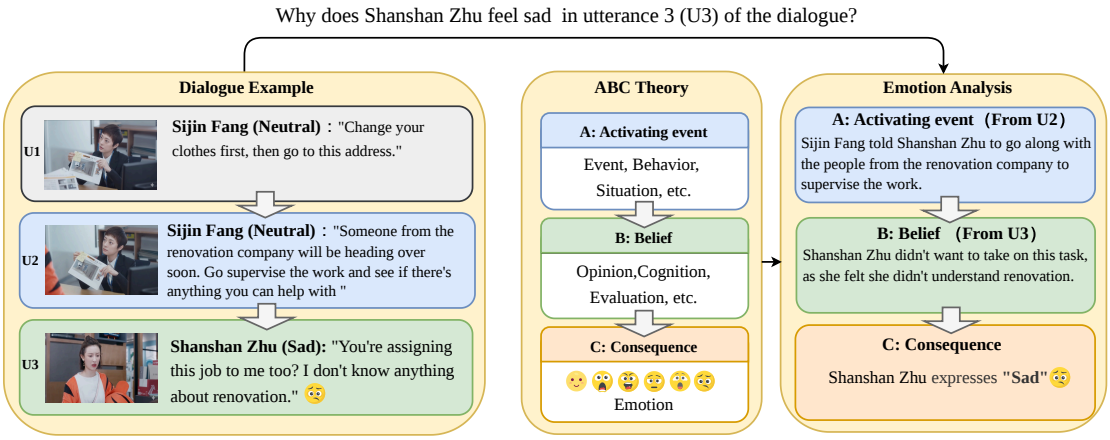

Figure 2: An example of the ABC annotation theory.

3.3 Annotation Process

Given the large amount of summary-style annotation required by our task, purely manual annotation is costly and inefficient, and often results in incomplete causal descriptions. To balance annotation quality and cost, we adopt a two-stage human–model collaborative annotation pipeline, as follows:

(1) Machine Pre-annotation. We employ Gemini-2.5-Pro (Comanici et al., 2025) to generate initial annotations using optimized prompts based on our guidelines (detailed in Appendix A.1). The machine-generated annotations include the positions of emotion cause utterances and multiple plausible emotion cause summaries expressed in different forms or from different perspectives.

(2) Human Check. Human annotators review and correct the inevitable errors in the machine-generated pre-annotations. Three undergraduate and graduate students perform the annotation, and one expert annotator resolves any inconsistencies. Annotators are compensated based on the quality and quantity of their work, and all undergo training and testing before formal annotation. To ensure data quality, we implement a double-verification workflow: each dialogue is independently checked by two annotators for both cause utterances and cause summaries. Consistent annotations are accepted directly, and disagreements are resolved by a third expert.

To better understand the necessity of the human check phase, we conducted a quantitative error analysis comparing the initial outputs of Gemini-2.5-Pro with the final human-verified annotations. The results indicate that the errors in the initial outputs primarily manifest in the following aspects: inaccurate or missing descriptions of the activating events and beliefs underlying the emotion causes, as well as insufficient identification of multimodal cues. Detailed results are provided in Appendix A.2.

To support the annotation process, we develop an annotation system, further details are provided in Appendix A.3. We provide a visual illustration of the full human–model collaborative annotation workflow in Appendix A.4.

Statistics	Train	Val	Test	Total
#Dialogues	546	78	157	781
#Utterances	7,126	977	2,033	10,136
w/ Cause	4,039	583	1,165	5,787
w/ Multi-cause	176	24	53	253
#Emotion-Cause Pairs	8,568	1,228	2,435	12,231
Avg. Utterances/Dialogue	13.05	12.52	12.94	12.97
Avg. Utterance Length	14.95	15.87	15.05	15.06
Avg. Summary Length	38.72	40.49	38.88	38.93

Table 1: Data statistics of MECESD. The summary length statistics are measured by the number of Chinese characters.

3.4 Data Statistics and Analysis

Overall Statistics. Table 1 presents the data statistics of MECESD, consisting of 781 dialogues and 10,136 utterances. Please note that we only annotate the cause of utterances that have non-neutral emotions, resulting in 5,787 utterances with extracted cause utterances and cause summaries, among which 253 utterances have more than one valid cause summary. We further examine the relative positions between emotion and cause utterances, observing that the majority of causes precede their corresponding emotions and occur

Dataset	Lang	Mod.	Source	#Utts	Cause-type		Multi-Ref.	Anno. Theory
					Extr.	Sum.		
ECF (Wang et al., 2023)	En	T,A,V	Friends TV	13,619	✓	×	×	-
MEC[4] (Wu et al., 2025)	Zh	T,A,V	Family comedy TVs	27468	✓	×	×	-
M [3]HG (Liang et al., 2025)	Zh	T,A,V	Multi-TVs	10,516	✓	×	×	-
ECGF (Wang et al., 2024)	En	T,A,V	Friends TV	13,619	×	✓	×	-
MECESD (Ours)	Zh	T,A,V	Multi-TVs	10,136	✓	✓	✓	ABC

Table 2: Comparison with existing datasets for MECAC. The columns denote Language (Lang), Modalities (Mod.), number of Utterances (#Utts). "Extr." and "Sum." denote Extraction and Summarization respectively. "Multi-Ref." stands for Multi-Reference.

within a short conversational distance (detailed in Appendix A.5). In addition, modality analysis reveals that more than 80% of emotion causes are conveyed through textual information, whereas fewer than 20% rely on visual or acoustic modalities, or require latent multimodal reasoning (detailed in Appendix A.6).

Inter-annotator Agreement. We measure inter-annotator agreement for emotion cause extraction using Cohen's kappa value (Cohen, 1960). The resulting Cohen's kappa value is 0.7511, which is higher than those reported for other emotion cause extraction datasets, such as ECF (0.6475) (Wang et al., 2023) and MECAD (0.6603) (Liang et al., 2025). These results indicate that the incorporation of the ABC theory and our human–model collaborative annotation strategy contributes to improved annotation quality and consistency.

Analysis of Cause Complexity Across Emotion Categories. We analyze cause complexity across different emotion categories using two metrics: the average length of cause summaries and the average number of cause utterances. As illustrated in Figure 3, the results show strong consistency between these metrics, highlighting the intrinsic correlation between the Multimodal Emotion Cause Extraction (MECE) and Multimodal Emotion Cause Summarization (MECS) sub-tasks. Negative emotions (e.g., Anger, Disgust, and Fear) exhibit higher causal complexity, as reflected by longer summaries and a larger number of associated utterances. These emotions are typically rooted in complex, multi-turn interactions or conflicts that require deeper contextual reasoning. In contrast, the causes of Happy and Surprise display lower complexity, as they are often triggered by single, sudden events with more direct causal relations.

Comparison with Existing Related Datasets. As shown in Table 2, we compare the MECESD

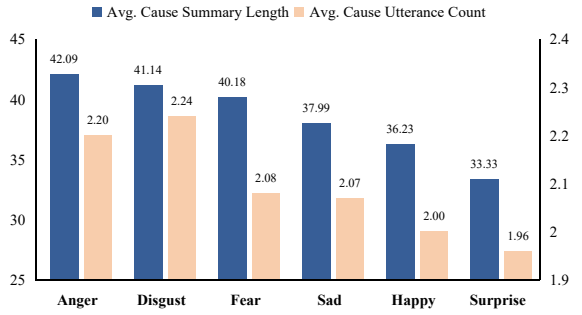

Figure 3: Distribution of average length of cause summary and average count of cause utterance. The cause summary length statistics are measured by the number of Chinese characters.

with existing MECAC datasets. To the best of our knowledge, MECESD is the first dataset jointly annotated for both extraction and summarization. In addition, unlike existing datasets that lack a specific theoretical annotation framework or use single-reference standards, MECESD is constructed based on a structured psychological theoretical framework and provides multi-dimensional cause summary annotations. Furthermore, while the only previous summarization dataset ECGF (Wang et al., 2024) is limited to a single source (Friends), MECESD ensures greater scenario diversity by sourcing from multiple TV series.

4 Method

4.1 Task Definition

Formally, given a dialogue $D = \{U_1, \ldots, U_n\}$ consisting of n chronologically ordered utterances. Each utterance U_i is represented as $U_i = [S_i, E_i, U_i^t, U_i^a, U_i^v]$, where S_i is the speaker of U_i, and E_i is its corresponding emotion label. U_i^t, U_i^a, and U_i^v denote the representations of the utterance from the textual, acoustic, and visual modalities, respectively. For a target utterance U_t $(1 \leq t \leq n)$, where its

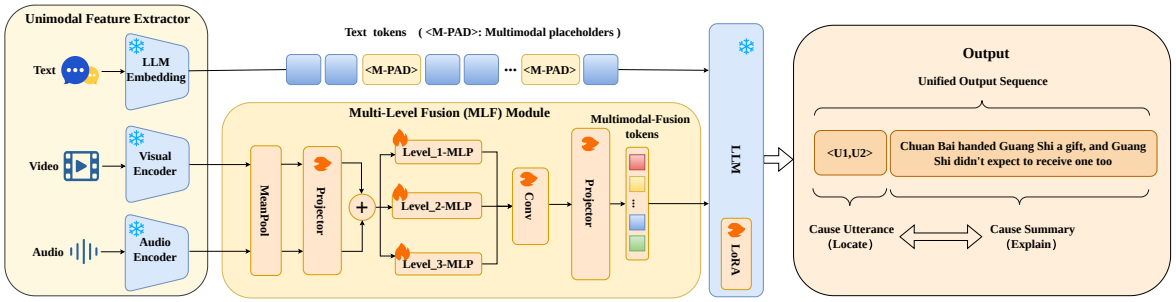

Figure 4: The overview of our proposed MPF-LLM framework.

speaker S_t exhibits a non-neutral emotion $E_t \in \{Happy, Surprise, Anger, Disgust, Fear, Sad\}$ (Ekman, 1992), the goal of the MECES task is to identify the set of cause utterances $C = \{U_j | j \in \{1, \ldots, n\}\}$ from the dialogue D that are responsible for E_t, and to generate a fluent and accurate natural language summary S' that describes the cause for the emotion E_t.

4.2 Overall Architecture

Figure 4 illustrates the Multimodal-Pre-Fusion-LLM (MPF-LLM) framework for the MECES task.

Unimodal Feature Extractor. For the text modality, we utilize the inherent text embedding layer of the Large Language Model (LLM) to encode the utterances and task instruction prompts for obtaining textual features (F_t) (please refer to the Appendix A.7 for specific prompts). For the visual and audio modalities, we employ CLIP ViT-L (Radford et al., 2021) and HuBERT-L (Hsu et al., 2021), respectively, to extract the corresponding visual features (F_v) and acoustic features (F_a).

Multi-Level Fusion Module. Non-textual modalities, such as acoustic features and visual cues, often carry crucial information for emotion causes in MECAC. Existing models, however, rely on the LLM to handle all multimodal fusion, which is challenging given that LLMs are primarily pre-trained on text. To address this, we propose a lightweight pre-fusion module, termed the Multi-Level Fusion (MLF) module, inspired by pre-fusion strategies in multimodal emotion recognition tasks (Yang et al., 2025b; Lian et al., 2025). The MLF module compresses and fuses audio-visual features of each utterance into a single pseudo-token, providing a compact, information-rich representation that allows the LLM to more effectively leverage non-textual emotional cues.

The detailed architecture of the MLF module is illustrated in Figure 4. First, for the raw video features F_v and audio features F_a of each utterance, we extract their global context via Mean Pooling. Subsequently, they are projected into a unified hidden space using modality-specific linear layers:

$$H_v = \text{Linear}_v(\text{MeanPool}(F_v)) \quad (1)$$

$$H_a = \text{Linear}_a(\text{MeanPool}(F_a)) \quad (2)$$

Then, we obtain a preliminary fused representation H_{av} via element-wise summation:

$$H_{av} = H_v + H_a \quad (3)$$

To capture inter-modal interactions at diverse abstraction levels, we employ three parallel Multi-Layer Perceptrons (MLPs) with varying hidden layer dimensions to process H_{av}:

$$H_i = \text{MLP}_i(H_{av}), \quad i \in \{1, 2, 3\} \quad (4)$$

Finally, to obtain a more compact multimodal representation, we use a 1×1 convolution layer to compress these features into a comprehensive fused representation H_f. This is followed by a linear projection layer, Linear_{out}, which maps H_f to an output dimension D_{out} that matches the input dimension of the LLM backbone:

$$H_f = \text{Conv}_{1\times1}([H_1; H_2; H_3]) \quad (5)$$

$$P_{out} = \text{Linear}_{out}(H_f) \quad (6)$$

The resulting P_{out} (pseudo-tokens) serve as the input to the subsequent LLM, replacing the corresponding Multimodal placeholders in the text tokens.

Model Training. During the training phase, we freeze the parameters of the LLM backbone and all unimodal encoders. Model fine-tuning is confined to the LoRA modules within the LLM, the Projector module and the MLF modules. We jointly model cause extraction and cause summarization as a unified sequence generation task. The model

is trained with a standard next-token prediction objective, optimizing by minimizing the negative log-likelihood loss. Furthermore, to balance the performance between the two sub-tasks, we experimented with multi-task learning loss functions. However, it does not yield performance improvements. For a given target sequence $Y = (y_1, y_2, \ldots, y_T)$ and model parameters θ, the loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}_{\text{MECES}}(\theta) = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, D; \theta) \quad (7)$$

where D represents the input context, $y_{<t}$ denotes the target tokens before time step t, and $P(y_t|\cdot)$ is the conditional probability of predicting the next token y_t.

5 Experiments

Data. We randomly split the newly annotated MECESD into train, val, and test sets at a 7/1/2 ratio. Data statistics are provided in Table 1.

Evaluation Metrics. Our MECES task comprises two subtasks: MECE and MECS. For the MECE task, similar to Wang et al. (2023), we adopt the weighted average F1 score as the metric. For the MECS task, we employ standard text generation metrics including BLEU-2, BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004), alongside semantic similarity-based metrics: BERTScore (Zhang et al., 2020) and Sentence-BERT (Reimers and Gurevych, 2019). Given that our task provides one or more references for each instance in the MECS task, we follow the standard practice for calculating all aforementioned summarization metrics. (Detailed information of metrics is provided in Appendix A.8). For experiments on the ECGF dataset, we adopt evaluation metrics consistent with Wang et al. (2024).

Compared Methods. Due to the limited research on MECE and MECS and the scarcity of models directly adaptable to our task, we select several methods for comparison on the MECES task, categorized by input modality as shown in Table 3. In the Text input setting, we include the SOTA textual emotion cause extraction model TSAM (Zhang et al., 2022), alongside LLMs such as Qwen2.5-7B (Yang et al., 2025a), ChatGLM3-6B (Du et al., 2022) and LLaMA2-7B (Touvron et al., 2023), as well as high-performing general-purpose models like Deepseek-V3.2 (DeepSeek-AI et al., 2025) and GPT-5 (OpenAI, 2025) operating on textual prompts. For the Multimodal (MM) input setting, we evaluate MLLMs including Qwen2.5-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025) against our proposed MPF-LLM. All models denoted with "(3-shot)" are evaluated via in-context learning, while the remaining models are fine-tuned on our dataset. Furthermore, for the MECS subtask, we conduct an additional comparison against previous methods such as ObG (Wang et al., 2024) on the ECGF dataset.

Implementation Details. We adopt ChatGLM3-6B-base as the LLM backbone, as it achieves superior performance compared to other backbone models (see Appendix A.9 for detailed results). In experiments on both datasets, models are trained exclusively on the training set and tuned on the validation set. We select the checkpoint achieving the best performance on the validation set for final evaluation on the test set and report these results. All experiments were conducted on 32GB Nvidia V100 and 40GB Nvidia A100 GPUs. (More detailed settings are provided in the Appendix A.10)

5.1 Main Results

Table 3 presents the main results of our proposed MPF-LLM and various methods on the MECESD dataset as benchmark results. Overall, our proposed MPF-LLM achieves the best performance across most metrics, demonstrating its effectiveness in jointly modeling emotion cause extraction and summarization with multimodal cues. Results on other datasets further confirm the robustness of our approach (see Appendix A.11).

Among the text-only methods, the smaller model TSAM yields relatively lower performance in cause extraction, while LLMs such as ChatGLM3-6B, LLaMA2-7B, and GPT-5 show better results. GPT-5 achieves strong F1 performance, reflecting its robust reasoning capability, but tends to favor abstractive generation. In contrast, DeepSeek-V3.2 shows competitive semantic alignment, as indicated by Sentence-BERT scores, while exhibiting comparatively lower precision in extractive localization.

We also evaluate multimodal LLMs with video input, Qwen2.5-VL-7B and InternVL3.5-2B. Despite supporting multimodal input, their performance lags behind text-only methods, likely due to challenges in maintaining long-range contextual modeling over extended conversational videos. This suggests that integrating multimodal informa-

Modality	Model	F1	BLEU-2	BLEU-4	METEOR	ROUGE-L	BERTScore	Sentence-BERT
Textual	TSAM	0.7331	/	/	/	/	/	/
	Qwen2.5-7B	0.7802	0.4582	0.3280	0.2824	0.4836	0.7866	0.7982
	ChatGLM3-6B	0.7855	0.4722	0.3503	0.2932	0.5125	0.7996	0.8055
	LLaMA2-7B	0.7820	0.4365	0.3075	0.2703	0.4662	0.7811	0.7790
	Deepseek-V3.2 (3-shot)	0.7235	0.3757	0.2640	**0.3162**	0.4732	0.7785	0.8059
	GPT-5 (3-shot)	0.7924	0.3193	0.1992	0.2739	0.4098	0.7518	0.7799
Multimodal	Qwen2.5-VL-7B (3-shot)	0.5135	0.3202	0.2014	0.2490	0.3788	0.7464	0.7435
	InternVL3.5-2B (3-shot)	0.4858	0.2351	0.1188	0.1922	0.2876	0.7032	0.6589
	MPF-LLM (Ours)	**0.7948**	**0.4754**	**0.3524**	0.2948	**0.5149**	**0.7998**	**0.8061**

Table 3: Main results on our MECESD. Models without the "(3-shot)" designation are fine-tuned on the training set.

tion in complex conversational tasks can introduce noise for models with limited parameter scales.

We further analyze the model performance from two perspectives: across emotion cause source types and across emotion categories. For emotion cause source types, the results show a clear trend: text-only causes are easiest to handle, followed by those conveyed through visual or audio cues, while latent causes requiring implicit reasoning are the most challenging. The experimental results indicate that all models exhibit a certain degree of performance variance across different emotion categories. This is likely tied to the inherent causal complexity of different emotions (as analyzed in section 3.4). Detailed results are provided in Appendix A.12 and Appendix A.13.

5.2 Analysis

Effectiveness of Joint-task Learning Framework. To validate the effectiveness of the joint-task learning framework, we compare the full MPF-LLM model with two single-task variants: MPF-LLM (Extr.), which only performs the MECE task, and MPF-LLM (Summary), which only performs the MECS task. As shown in Table 4, the jointly trained MPF-LLM outperforms both single-task variants across all metrics, demonstrating that joint training effectively leverages complementary information between the two tasks to enhance context understanding and overall performance. We further analyze the correlation between the extraction and summarization sub-tasks. The results indicate that MPF-LLM maintains high internal consistency in causal reasoning, rather than treating the sub-tasks independently. Detailed analyses are provided in Appendix A.14.

Effects of Different Modalities. We examine the impact of removing different modalities on MPF-LLM's performance. Results show that removing either the visual (w/o V) or audio (w/o A) modality leads to performance drops across most metrics, suggesting that both modalities provide important cues for emotion cause analysis. The drop is larger when removing visual information, indicating that visual information plays a more crucial role in MECES task. Removing both modalities (w/o A,V) results in the lowest performance in emotion cause extraction, highlighting the necessity of multimodal information for accurate emotion cause localization.

Effectiveness of the MLF Module. To evaluate the contribution of the MLF module, we replace it with two baseline strategies: 1) Respective Pseudo-tokens (RP), which feeds average-pooled visual and audio features as separate tokens; 2) Sum, which sums the average-pooled visual and audio features and feeds the result directly into the LLM. Results show that RP leads to a notable performance decline. While Sum achieves slightly better results on BLEU-4, our full model consistently outperforms it on semantic metrics (BERTScore, Sentence-BERT). These results confirm the effectiveness of the MLF module in efficiently fusing audio and visual cues, enabling the LLM to better leverage non-textual information for more accurate emotion cause extraction and summarization.

Human Evaluation of Emotion Cause Summary Quality. Beyond evaluating the overall similarity between the generated cause summaries and the references, we further conduct a human evaluation to analyze the quality of the generated summaries through the perspective of the ABC theory, which serves as the core theoretical foundation for the construction of the MECESD. We randomly sampled 50 instances from the test set to conduct a human evaluation of MPF-LLM against two strong baselines: ChatGLM3-6B and DeepSeek-V3.2. Specifically, based on the ABC theory, hu-

	F1	BLEU-2	BLEU-4	METEOR	ROUGE-L	BERTScore	Sentence-BERT
w/o V	0.7896	0.4724	0.3504	0.2926	0.5126	0.7987	0.8045
w/o A	0.7922	0.4744	**0.3530**	0.2943	0.5126	0.7988	0.8052
w/o A,V	0.7855	0.4722	0.3503	0.2932	0.5125	0.7996	0.8055
w/o MLF (RP)	0.7920	0.4694	0.3485	0.2916	0.5121	0.7981	0.8054
w/o MLF (Sum)	0.7925	0.4749	0.3528	0.2938	0.5118	0.7982	0.8053
MPF-LLM (Extr.)	0.7879	/	/	/	/	/	/
MPF-LLM (Summary)	/	0.4637	0.3432	0.2889	0.5097	0.7983	0.8053
MPF-LLM	**0.7948**	**0.4754**	0.3524	**0.2948**	**0.5149**	**0.7998**	**0.8061**

Table 4: Ablation study of different components on the MECESD.

Model	A Score	B Score	Overall Score
ChatGLM3-6B	4.20	3.33	3.24
DeepSeek-V3.2 (3-shot)	4.10	3.30	3.16
MPF-LLM	**4.40**	**3.40**	**3.40**

Table 5: Human evaluation of emotion cause summary.

man evaluators scored each generated causal summary on a scale of 0 to 5 across three dimensions: (1) Activating event (A), measuring the correctness and completeness of the described activating event; (2) Belief (B), measuring the correctness and plausibility of the speaker's beliefs; and (3) Overall Score, evaluating whether the generated summary correctly integrates both the activating event and the belief into a coherent and rational summary. As shown in Table 5, the results demonstrate that MPF-LLM consistently outperforms both baselines across all three dimensions, indicating its superior capability in generating emotion cause summaries that better align with the ABC theory. Furthermore, this observation is consistent with the trends exhibited by the automatic evaluation metrics, demonstrating that our automatic evaluation metrics can effectively reflect the models' ability in capturing emotional causal logic. Further analysis reveals that all models score significantly higher on the Activating event (A) dimension than on the Belief (B) dimension. This discrepancy suggests that capturing the speaker's subjective beliefs regarding emotion causes remains a shared challenge for current models, highlighting a critical direction for future improvements in emotion cause analysis.

6 Conclusion

In this work, we introduce the MECES task, which aims to simultaneously locate and explain emotion causes in conversations, addressing the limitations of treating extraction and generation in isolation. We construct MECESD, a high-quality dataset annotated under the guidance of the ABC theory, enabling comprehensive and consistent emotion cause analysis. Furthermore, we propose MPF-LLM, an end-to-end framework equipped with a Multi-Level Fusion module that effectively aligns non-textual cues with the LLM backbone to enhance multimodal understanding. Through joint learning of extraction and summarization, our approach demonstrates that the two tasks provide complementary information, leading to superior performance and more interpretable emotion cause analysis.

Limitations

First, although we adopt a human–model collaborative annotation strategy to balance annotation quality and cost, the process remains labor-intensive and time-consuming. As a result, the current dataset scale is relatively limited. In future work, we plan to explore more efficient and scalable annotation paradigms to further expand datasets in this domain.

Second, the model proposed in this paper is intended as a preliminary baseline. Its performance remains limited when handling complex multimodal emotion causes and latent causes that require deep reasoning. Future research will focus on developing more advanced model architectures and reasoning mechanisms to better address these challenges.

Ethical Considerations

This study does not involve the collection of new raw data. All annotation work was conducted on the publicly available M3ED (Zhao et al., 2022) dataset, and permission for its use has been obtained from the original dataset owners. We have carefully screened the data to ensure it contains no harmful or offensive information. Additionally,

all annotators were compensated with wages corresponding to their workload. Furthermore, we will place restrictions on the use of our dataset, requiring researchers to commit to using the data in a responsible manner. We explicitly stipulate that the dataset is strictly limited to non-commercial academic research purposes. Any commercial or other non-academic use is strictly prohibited.

Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Natural Science Foundation of China (Grant No. 62306202 and 62476187), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Albert Ellis. 1957. Rational psychotherapy and individual psychology. *Journal of individual psychology*, 13(1):38.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71, Bangkok, Thailand. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. 2024. UniMEEC: Towards unified multimodal emotion recognition and emotion cause. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5248–5261, Miami, Florida, USA. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Trans. Multimedia Comput. Commun. Appl.* Just Accepted.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *Preprint*, arXiv:2501.16566.

Qiao Liang, Ying Shen, Tiantian Chen, and Lin Zhang. 2025. M3HG: Multimodal, multi-scale, and multi-type node heterogeneous graph for emotion cause triplet extraction in conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11416–11431, Vienna, Austria. Association for Computational Linguistics.

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

OpenAI. 2025. Introducing GPT-5. Accessed: 2025-12-18.

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*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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, 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*, MM '24, page 5820–5828, New York, NY, USA. Association for Computing Machinery.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Dan Wu, Xincheng Ju, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. 2025. Emotion across modalities and cultures: Multilingual multimodal emotion-cause analysis with memory-inspired framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 5775–5783, New York, NY, USA. Association for Computing Machinery.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025a. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2025b. Mse-adapter: a lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Gemini Model Pre-annotation Prompts

As shown in Table 6, prompts optimized based on our annotation guidelines were used for Gemini's initial annotation. Furthermore, the 3-shot examples used in the experiments were also refined based on this prompt.

A.2 Error Analysis of Machine Pre-annotations

To further investigate the potential error patterns during the machine pre-annotation stage, we further conduct a quantitative error analysis of the machine pre-annotations. Specifically, we randomly sample 50 instances and compare the initial outputs produced by Gemini-2.5-Pro with the final human-verified annotations. We categorize the major error patterns into the following types.

- **Inaccurate or Missing Activating events (A):** The model fails to accurately capture or completely omits the objective events, behaviors, or situations that trigger the emotion, accounting for 12%.
- **Inaccurate or Missing Beliefs (B):** The model overlooks or misidentifies the speaker's subjective opinions, cognitions, or evaluations regarding the event, accounting for 20%.
- **Missing Multimodal Cues:** The model over-relies on textual information, missing critical causal clues conveyed through visual or acoustic modalities, accounting for 8%.

Additionally, we also observe errors in third-person pronoun conversion and ungrounded causal reasoning (ignoring source dialogue facts) in the pre-annotations. Despite the aforementioned errors in the machine pre-annotations, the subsequent human verification mechanism effectively safeguards the data quality. Our inspection of the final annotations for these 50 sampled instances confirms that all identified issues have been successfully rectified.

A.3 Annotation Tool

We show the annotation interface of our annotation tool in Figure 5.

A.4 Annotation Process

We show our "human-model collaborative" annotation pipeline in Figure 6.

A.5 Relative Position of Emotion and Cause Utterances

We further analyze the distribution of emotion causes. Figure 7 illustrates the relative position between emotion utterances and their corresponding cause utterances. Specifically, the horizontal axis represents the relative distance, calculated by subtracting the index of the emotion utterance from the index of the cause utterance. As depicted in the figure, cause utterances are highly concentrated near corresponding emotion utterances, particularly in preceding positions, consistent with the common "cause-before-effect" logic commonly found in natural language. About 41.44% of the causes are located within the same utterance as the emotion, suggesting that emotions and their causes are frequently co-expressed. Moreover, approximately 4.03% of "backward causes" (i.e., the cause utterance appears after the emotion utterance). This highlights the necessity of analyzing the entire dialogue rather than relying solely on historical context.

A.6 Distribution of Emotion Cause Source Types

We performed an analysis of the source types for emotion causes in our dataset. As shown in Figure 8. The text modality serves as the predominant source for emotion causes. In addition, approximately 12.2% of emotion causes are directly derived from the visual or audio modalities. More challengingly, 7.4% of the causes in the dataset are classified as "Latent Causes". These causes require complex reasoning over the multimodal context to identify, such as the comprehension of implicit meanings or the application of external knowledge. This distribution underscores the diversity of our dataset and its alignment with the complexity of real-world emotion antecedents.

A.7 Prompt Template for MPF-LLM Input

As shown in Table 7, we organize the input to MPF-LLM in a unified prompt format, including four components: Instruction, Dialogue Context, Target Emotion Utterance, and Output.

A.8 Detailed Information of Metrics

For the MECE task, similar to (Wang et al., 2023), we adopt the weighted average F1 score as the metric. For the MECS task, we introduce multiple metrics tailored to the specific characteristics of

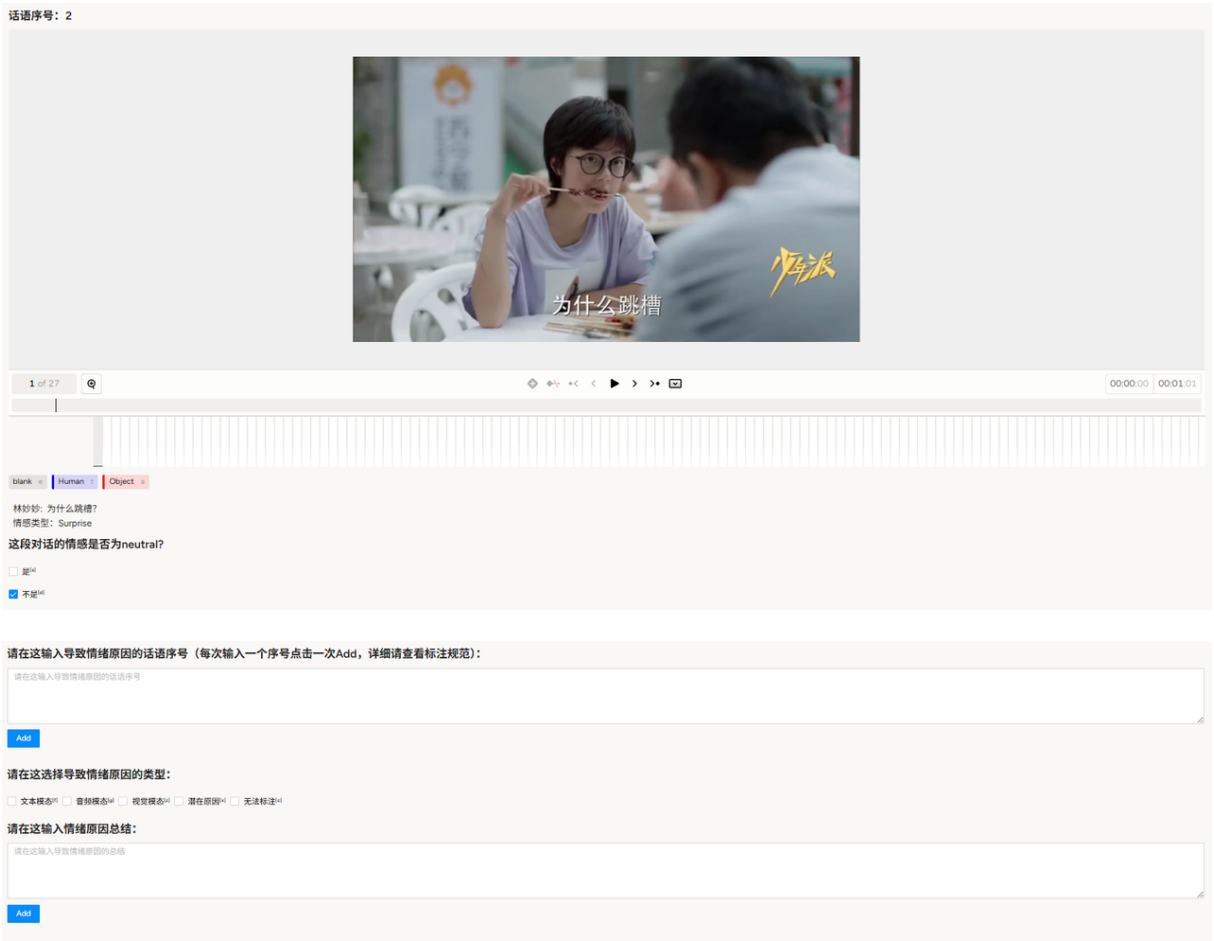

Figure 5: Annotation interface of our annotation tool.

	Prompt Content
Instruction	You are an expert specialized in dialogue emotion cause analysis. Please answer according to the following task definition and examples.
Task Definition	I will provide a **multimodal dialogue and the corresponding video**, along with a specific "Target Utterance". Based on the dialogue context, analyze the cause behind the speaker's emotion in the target utterance. Specifically, you need to analyze the emotion triggering process based on the definition of emotion causes in the **ABC Theory of Emotion** (where emotion causes consist of *Triggering Events* [i.e., objective events, behaviors, situations] and *Beliefs* [i.e., viewpoints, cognitions, evaluations]). Complete the following two sub-tasks: 1. **Locate Emotion Cause**: Identify the utterance indices that led to the target emotion. 2. **Generate Cause Summary**: Generate a concise explanation of the emotion cause.
Example 1	**Input:** { U_1_Neutral. Zhong Xiaoqin: Did that woman come to find you again? That must mean Liang Zhengxian settled things with her. ... U_3_Neutral. Wang Manni: Liang Zhengxian said he wants to buy a house in Shanghai and settle down here. ... U_4_Surprise. Zhong Xiaoqin: That's very sincere! ... **Target:** U_4_Surprise } **Output:** { Cause Indices: U_3, U_4; Cause Summary: Wang Manni said Liang wants to buy a house in Shanghai and settle down, and Zhong Xiaoqin thinks this shows Liang's sincerity. }
Example 2	**Input:** { U_1_Neutral. Fang Sijin: Change your clothes first, then go to this address. U_2_Neutral. Fang Sijin: People from the renovation company will be there soon. Go supervise... U_3_Sad. Zhu Shanshan: Do I have to do this job too? Renovation, I don't understand it. ... **Target:** U_3_Sad } **Output:** { Cause Indices: U_2, U_3; Cause Summary: Fang Sijin asked Zhu Shanshan to go supervise the renovation team, but Zhu didn't want to do it because she felt she didn't understand renovation. }
Example 3	**Input:** { U_1_Neutral. Lin Miaomiao: My request is that I want to ride a bike to school. ... U_6_Disgust. Lin Dawei: Your skills are too poor. ... **Target:** U_6_Disgust } **Output:** { Cause Indices: U_1, U_6; Cause Summary: Lin Miaomiao proposed riding a bike to school, but Lin Dawei thought her skills were too poor and the request was unrealistic. }
Format	* Emotion cause utterance indices should be sorted in ascending order and separated by commas. * The emotion cause summary should follow the style of the examples, focusing on the content provided in the dialogue, and be concise, objective, and cover key contextual information.
User Input	**Input: .....**

Table 6: Gemini Model Pre-annotation Prompts.

the task. We first report a set of widely-used standard text generation metrics, including BLEU-2, BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004) to measure the lexical overlap between generated summaries and references. Secondly, we posit that the core evaluation for emotion cause summary lies in determining whether the generated summary accurately captures the main emotion cause. Therefore, we further introduce semantic similarity-based metrics: BERTscore (Zhang et al., 2020) and Sentence-BERT (Reimers and Gurevych, 2019). For BLEU-2, BLEU-4, METEOR, and ROUGE-L, we utilize the nlg-eval (Sharma et al., 2017) toolkit. For BERTScore, we adopt the standard bert-base-chinese model for evaluation.

A.9 Impact of Model Configurations

To further investigate the factors that may significantly affect model performance, we conduct additional experiments in this section. To facilitate a more concise and holistic comparison across different window settings, we aggregate the fine-grained evaluation metrics into three macro-indicators: F1, lexical overlap (the arithmetic mean of n-gram based metrics, including BLEU-2, BLEU-4, METEOR, and ROUGE-L), and semantic similarity (the embedding-based metrics, including BERTScore and Sentence-BERT). The results are presented in Table 8.

The Effect of Context Window Size. To investigate the impact of context window size on model performance, we conducted experiments with varying window size configurations. We define the target utterance as the anchor position and denote $window(m, n)$ as the context window encompassing the m preceding and n succeeding utterances. Specifically, $window(max, max)$ denotes the full dialogue context. This paper adopts $window(8, 3)$ as the default experimental setting.

Experimental results indicate that for the MECE task, smaller context windows yield superior per-

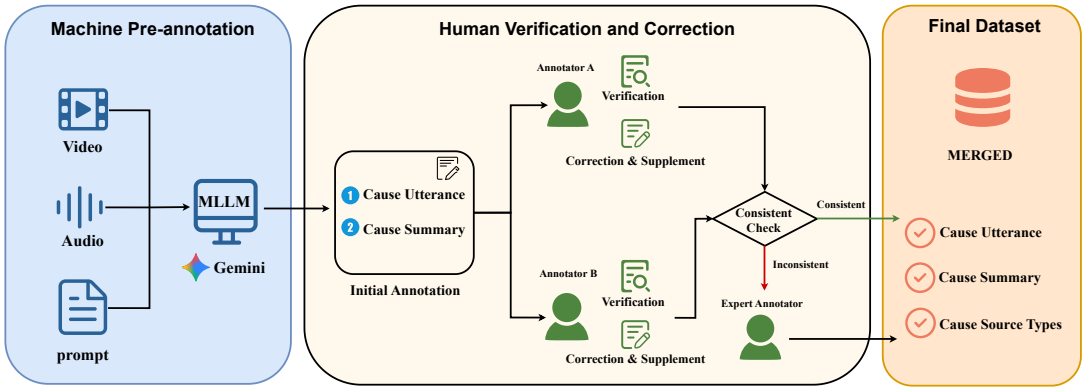

Figure 6: Annotation process.

	Prompt Template for MPF-LLM Input
Instruction	You are an expert in emotion analysis and emotion cause extraction. I will provide a "dialogue context" involving two speakers with corresponding emotion labels and multimodal information. Your task is to identify the "cause utterances" that trigger the specific emotion in the "target emotion utterance," output the list of indices for these cause utterances, and generate a concise emotion-cause summary.
Dialogue Context	U1_Anger. Situ Mo: Then kick me! <multimodal placeholders> U2_Neutral. Gu Weiyi: I dare not. <multimodal placeholders> U3_Anger. Situ Mo: What do you mean you dare not? You even dared to let someone else tie your tie for you. <multimodal placeholders> U4_Anger. Gu Weiyi: She ambushed me; I am innocent. <multimodal placeholders> U5_Anger. Situ Mo: Really? You looked like you enjoyed it. <multimodal placeholders> U6_Anger. Situ Mo: See! I hit the mark and now you have nothing to say! <multimodal placeholders>
Target Emotion Utterance	U3_Anger. Situ Mo: What do you mean you dare not? You even dared to let someone else tie your tie for you. <multimodal placeholders>
Output	[U3], Gu Weiyi let another girl tie his tie, and Situ Mo felt that Gu Weiyi actually dared to do such a thing.

Table 7: Example of the unified prompt template used by MPF-LLM for the MECES task. The symbol <multimodal placeholders> denotes the positions where fused multimodal pseudo-tokens are inserted.

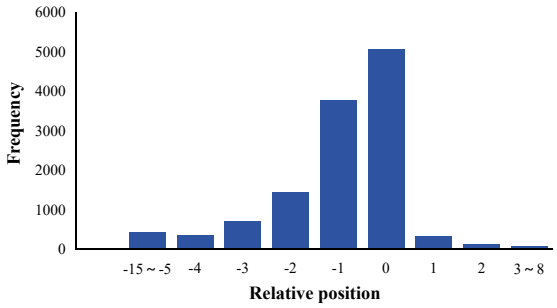

Figure 7: Relative position of emotion and causes

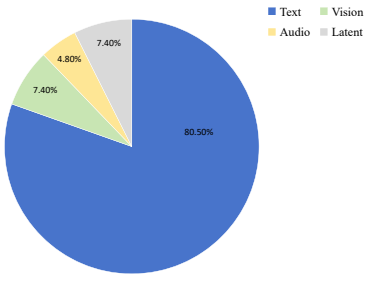

Figure 8: Distribution of emotion cause source types

formance. Specifically, the model achieves the highest F1 score under the $window(3, 1)$ setting. Conversely, utilizing the full context, denoted as $window(max, max)$, leads to a significant degradation in the F1 metric. This observation aligns with the analysis of the "relative position of emotion and cause utterances" presented in Appendix A.5, given that most causes are distributed in the

Setting	F1	Lex.	Sem.
Context Window Size			
Window (3, 1)	**0.7967**	0.4042	0.7993
Window (4, 2)	0.7937	0.4021	0.8008
Window (8, 3)	0.7948	**0.4093**	**0.8029**
Window (11, 5)	0.7874	0.4033	0.8000
Window (15, 8)	0.7891	0.4021	0.8015
Window (max, max)	0.7816	0.3993	0.8005
LLM Backbone			
LLaMA2-7B	0.7794	0.3716	0.7813
Qwen2.5-7B	0.7900	0.4029	0.8005
Qwen2.5-3B	0.7711	0.3749	0.7887
ChatGLM3-6B	**0.7948**	**0.4093**	**0.8029**

Table 8: Performance comparison across different context window sizes and LLM backbones. "Lex." and "Sem." denote lexical overlap metrics and semantic similarity metrics, respectively.

immediate vicinity of the target utterance. The smaller context windows enable the model to more accurately localize the cause.

In the MECS task, we observe that $window(8, 3)$ demonstrates the best overall performance, achieving the highest scores in both lexical overlap and semantic similarity. However, further reducing or expanding the window size yields no performance gains. These findings suggest the MECS task requires a moderate context window to provide necessary conversational background (e.g., interpersonal relationships and event context) for facilitating emotion cause summarization. Conversely, an excessively long context may introduce irrelevant noise, thereby degrading model performance.

The Effect of LLM Backbone. To investigate the the impact across different base models, we conducted evaluations on Llama2-7B, Qwen2.5-7B, and ChatGLM3-6B, with ChatGLM3-6B serving as the default backbone in this work. Experimental results demonstrate that ChatGLM3-6B outperforms other baselines across all metrics, followed by Qwen2.5-7B, which also exhibits competitive performance. In contrast, Llama2-7B and Qwen2.5-3B show a significant performance gap. We attribute the suboptimal performance of Llama2-7B to its English-centric pre-training corpus, which constrains its capabilities in Chinese text understanding and generation. Meanwhile, the performance decline of Qwen2.5-3B is primarily ascribed to its limited parameter size. Furthermore, compared to the text-only Qwen2.5-7B and ChatGLM3-6B baselines presented in Tables 3 and Table 8, our method achieves substantial improvements. This further validates the effectiveness and robustness of our method across different base models.

A.10 Hyperparameter Configuration

We present the detailed hyperparameter settings used for fine-tuning in Table 9. The model was trained using the LoRA (Low-Rank Adaptation) technique.

Hyperparameter	Value
Training Setup	
Number of GPUs	2
Number of Epochs	3
Learning Rate	0.0001
Batch Size (per device)	1
Gradient Accumulation Steps	4
Effective Batch Size	8
LoRA Configuration	
LoRA Rank (r)	8
LoRA Alpha (α)	32
LoRA Dropout	0.1

Table 9: Hyperparameters used for model training.

A.11 Results of the MECS task on the ECGF Dataset

Table 10 presents the comparison results of our MPF-LLM method and baseline models such as ObG on the MECS task for the ECGF dataset. Given that the original ECGF dataset lacks explicit annotations for MECE, we employed a data augmentation strategy to construct a corpus suitable for joint training, analogous to the MECES task. Specifically, we utilized the Qwen2.5-7B model, guided by meticulously crafted prompts, to reverse-infer and generate corresponding emotion cause utterance indices from the gold-standard emotion cause summaries. These generated utterance indices were subsequently used as pseudo-labels for the MECE task. Experimental results demonstrate that this joint training approach, enabled by data augmentation, yields significant performance improvements. This further suggests a strong correlation between the two subtasks.

Overall, MPF-LLM (Data Augmentation) demonstrates superior performance across most key metrics. Compared to ObG, MPF-LLM (Data Augmentation) achieves notable improvements on most

	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	F_BERT
Others	Gemini-Pro-Vision (3-shot)	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
	Flan-ObG	0.4967	0.4313	0.3924	0.3631	0.3042	0.4781	3.0808	0.7711
Ours	MPF-LLM	0.4864	0.4154	0.3743	0.3455	0.2991	0.4848	3.0226	0.7841
	MPF-LLM (Data Augmentation)	**0.5096**	**0.4386**	**0.3967**	**0.3663**	**0.3095**	**0.4922**	**3.1056**	**0.7870**

Table 10: Results of the MECS task on the ECGF dataset. Baseline results are cited from original paper (Wang et al., 2024). MPF-LLM (Data Augmentation) denotes the MPF-LLM model incorporating data augmentation strategies.

major evaluation metrics, particularly ROUGE-L and BERTScore. Although it scores slightly lower on CIDEr, its leading advantage in BERTScore and ROUGE-L suggests that our model possesses a superior capability for semantic comprehension and the summarization of core emotion causes. This is particularly critical for the MECS task, which is highly dependent on understanding core emotion cause.

A.12 Performance Analysis across Emotion Cause Source Types

To further investigate the capabilities of the models in processing different types of emotion causes, we stratified the samples in the test set into three categories based on the source of the cause: Text Only, Visual and Audio, and Latent. This facilitated a fine-grained evaluation. We benchmarked MPF-LLM against representative baseline models (ChatGLM3-6B, Deepseek-V3.2, and GPT-5), with detailed experimental results presented in the Table 11.

Initially, the results reveal significant performance disparities across different emotion cause types, following a general trend of Text Only > Visual and Audio > Latent. Specifically, the superior performance of MPF-LLM in the dominant category (Text Only) demonstrates that its architecture effectively leverages multimodal context to enhance textual reasoning. However, despite being designed to better utilize multimodal information, we observed that MPF-LLM underperforms compared to text-only models (ChatGLM3-6B) and GPT-5 in the Multimodal cause category. We attribute this primarily to the nature of real-world dialogue data, where textual content often embeds strong implicit multimodal cues (e.g., the utterance "Wow, is this gift for me?" allows for the easy inference of a "handing over a gift" scenario). This enables models to infer causes solely through textual backtracking, without necessitating direct reliance on visual or acoustic information.

	Model	F1	Lex.	Sem.
Text Only	ChatGLM3-6B	0.7947	0.4911	0.8090
	Deepseek-V3.2 (3-shot)	0.7350	0.4411	0.8003
	GPT-5 (3-shot)	0.8061	0.4097	0.7720
	MPF-LLM	**0.8077**	**0.4944**	**0.8098**
Visual and Audio	ChatGLM3-6B	0.7722	**0.4904**	**0.7683**
	Deepseek-V3.2 (3-shot)	0.6978	0.3991	0.7402
	GPT-5 (3-shot)	**0.8003**	0.3674	0.7242
	MPF-LLM	0.7510	0.4644	0.7628
Latent	ChatGLM3-6B	**0.6969**	0.4131	0.7687
	Deepseek-V3.2 (3-shot)	0.6168	0.3575	0.7559
	GPT-5 (3-shot)	0.6446	0.3315	0.7358
	MPF-LLM	0.6825	**0.4180**	**0.7700**
All types	ChatGLM3-6B	0.7855	0.4070	0.8025
	Deepseek-V3.2 (3-shot)	0.7235	0.3572	0.7922
	GPT-5 (3-shot)	0.7924	0.3005	0.7658
	MPF-LLM	**0.7948**	**0.4093**	**0.8029**

Table 11: Performance comparison across different Emotion Cause Source Type. "Lex." and "Sem." denote Lexical Overlap metrics and Semantic Similarity metrics, respectively.

Furthermore, given the scarcity of samples in the multimodal category, MPF-LLM may have overfitted to the dominant textual modality logic. In such long-tail scenarios, the MLF module may not have achieved optimal feature extraction and fusion states. Consequently, when processing complex explicit visual or acoustic signals, its stability lags behind that of ChatGLM3-6B, which relies solely on textual inference. Furthermore, in the most challenging latent category, all models exhibit a significant decline in performance, further corroborating the immense challenge of deep contextual reasoning in the absence of explicit cues.

A.13 Performance Analysis across Emotion Categories

We conduct a fine-grained evaluation on the test set by splitting instances into six emotion categories: Happy, Surprise, Anger, Disgust, Fear, and Sad. We compare our MPF-LLM with two representative LLM baselines, ChatGLM3-6B and GPT-5. Following the setting in Appendix A.9, we report three macro indicators: F1, lexical overlap (Lex.), and semantic similarity (Sem.). The results are

Metric	Model	Happy	Surprise	Anger	Disgust	Fear	Sad
F1	MPF-LLM	0.7682	**0.8649**	0.7915	0.8400	0.7573	0.7802
	ChatGLM3-6B	0.7594	**0.8582**	0.7864	0.8182	0.7723	0.7579
	GPT-5 (3-shot)	0.7754	**0.8885**	0.7751	0.8477	0.7097	0.7689
Lex.	MPF-LLM	0.3767	**0.4755**	0.4101	0.4246	0.2994	0.4082
	ChatGLM3-6B	0.3765	**0.4794**	0.3940	0.4109	0.3157	0.4124
	GPT-5 (3-shot)	0.2760	**0.3411**	0.3018	0.3025	0.2575	0.3065
Sem.	MPF-LLM	0.7779	**0.8289**	0.8100	0.8116	0.7644	0.8058
	ChatGLM3-6B	0.7775	**0.8290**	0.8089	0.8031	0.7731	0.8067
	GPT-5 (3-shot)	0.7396	**0.7862**	0.7703	0.7706	0.7640	0.7736

Table 12: Performance comparison across different emotion categories.

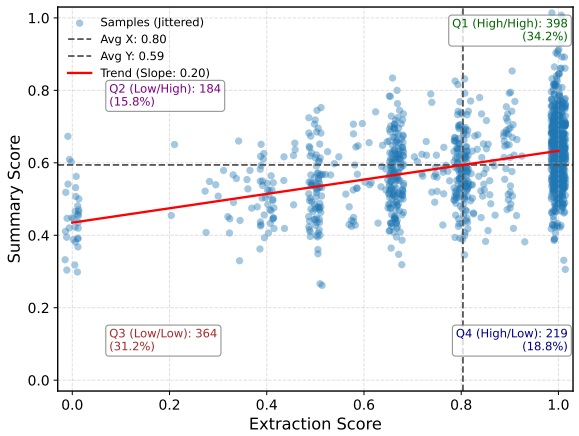

Figure 9: Instance-level performance correlation analysis between MECE and MECS tasks.

shown in Table 12.

The experimental results indicate that all models exhibit a certain degree of performance variance across different emotion categories. This is likely tied to the inherent causal complexity of different emotions (as analyzed in Section 3.4). For instance, models generally struggle more with "Fear" due to its often implicit and complex causes, while achieving optimal results on "Surprise", which tends to have more explicit triggers.

A.14 Fine-grained Analysis of Model Performance on the MECE and MECS Task

To investigate the fine-grained performance of MPF-LLM on the joint task of MECE and MECS, we conducted a visual analysis of the model's results on the test set. As illustrated in Figure 9, the horizontal and vertical axes of the scatter plot correspond to the extraction score and summary score of each sample, respectively.

Statistical analysis reveals a significant positive correlation between the model's performance on these two metrics (Pearson $r = 0.4336$, regression slope $k = 0.20$). The sample distribution exhibits a distinct diagonal clustering pattern, with the first (Q1: High/High) and third (Q3: Low/Low) quadrants accounting for 65.4% of the total samples. This distribution suggests a high degree of internal consistency between the model's capabilities to "Locate" and "Explain" emotional causes.

Specifically, the high density in Q1 indicates that the model achieves alignment between localization and explanation in most cases. Conversely, the clustering in Q3 implies that when the model fails to capture the emotional logic, both capabilities degrade synchronously. This reflects that the model performs unified reasoning rather than treating the tasks in isolation. Furthermore, outliers in Q2 (15.8%) and Q4 (18.8%) highlight reasoning biases in specific scenarios. The phenomenon of "accurate localization but deviant explanation" in Q4 suggests the model may be merely fitting positional features without genuinely comprehending the emotional cause. Meanwhile, the "accurate explanation but deviant localization" observed in Q2 reflects a grounding failure, where the generated content lacks support from specific contextual evidence, indicating flaws in the reasoning process.

In summary, the experimental evidence verifies that MPF-LLM possesses high internal consistency in causal reasoning. This demonstrates that the model is not merely fitting the objective functions of individual tasks, but is attempting to construct a comprehensive analytical framework spanning from localization to explanation. Moreover, the existence of the phenomena in Q2 and Q4 further validates the necessity of the proposed MECES framework for building an interpretable and robust emotion analysis system.