JX4MEI: Multimodal Semantically-Enhanced LLM for Joint Multimodal Emotion-Intent Explanation and Classification

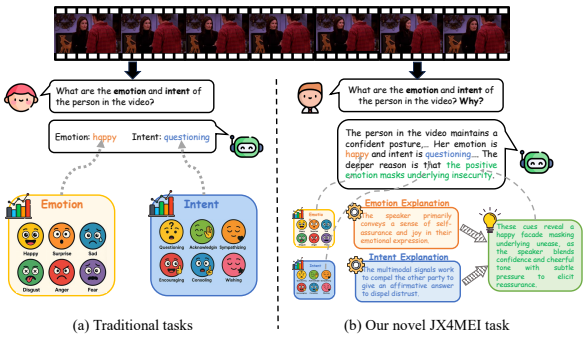
Yijie Huang*, **Xiaocui Yang***, **Shi Feng**†, **Daling Wang, Yifei Zhang, Ning Yuan, Zhuoyue Jia, Wen Zhang**
School of Computer Science and Engineering, Northeastern University
Shenyang 110819, China
{fengshi, yangxiaocui, wangdaling, zhangyifei}@cse.neu.edu.cn
{2401837, 2401954, 2401845, 2401967}@stu.neu.edu.cn

Abstract

Existing multimodal emotion and intent recognition tasks predominantly focus on classification, overlooking the underlying rationale and intrinsic connections between these states. Bridging this gap, we propose **Joint Multimodal Emotion-Intent Explanation and Classification**, **JX4MEI**, a novel task requiring the model to jointly predict emotion and intent, while generating natural language explanations for why they co-occur. To support this, we present **XMEI-dataset**, a large-scale benchmark of 15,461 multimodal samples covering 7 emotion and 9 intent categories across text, audio, and visual modalities. Unlike prior works, our dataset provides fine-grained rationales for emotion, intent, and their causal interplay, curated via a rigorous pipeline involving Chain-of-Thought generation and strict human refinement to eliminate model artifacts. Furthermore, we propose **XMEI-Qwen**, a model equipped with a novel **Language-Query Former (LQ-Former)**. By leveraging modality-specific captions as semantic queries, LQ-Former injects explicit semantic guidance into feature alignment, significantly enhancing reasoning capabilities. Empirical experiments demonstrate that XMEI-Qwen sets a new state-of-the-art on the JX4MEI task, outperforming competitive baselines in both prediction and explanation generation. Code: https://github.com/OrangeYeah1027/JX4MEI.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) (Wang et al., 2025b; Yang et al., 2024; Li et al., 2024; Zhang et al., 2024) have accelerated the paradigm shift in sentiment analysis towards multimodal data (Lee et al., 2025; Xu et al., 2025b; Das and Singh, 2023). Nevertheless, most existing systems still treat emotion recognition and intent detection as separate tasks (Xu et al., 2025a; Lian et al., 2023a; Zhang et al., 2022). This separation is inherently limited because it ignores the intrinsic link between emotion and intent. As Minsky observed (Minsky, 1986), emotions are not merely affective states but powerful communicators of our intentions. Therefore, joint modeling of emotion and intent is both theoretically and practically essential.



(a) Traditional tasks (b) Our novel JX4MEI task

Figure 1: An illustration of our proposed JX4MEI task in comparison to traditional approaches.

To this end, Liu et al. (2024) introduces the Multimodal Emotion and Intent Joint Understanding (MEIJU) task. However, this *classification-first* paradigm, also adopted by similar approaches (Singh et al., 2022), is inherently limited. By focusing only on identifying *what* the emotion and intent are, it overlooks the more critical question of *why*. Specifically, these models fail to elucidate the distinct multimodal causal factors and the reasoning linking emotion to intent. While identifying *what* users feel is foundational, explaining why bridges the gap to trustworthy Artificial Intelligence (DW, 2019). In high-stakes application scenarios such as mental health counseling (Sharma and De Choudhury, 2018), systems must decipher the causal interplay to formulate appropriate responses, for instance, distinguishing whether a user's anger stems from frustration (intent to complain) or fear (in-

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20242–20261

tent to seek help). To bridge this gap, we propose **Joint Multimodal Emotion-Intent Explanation and Classification (JX4MEI)**. This task extends beyond simple label prediction to causal reasoning, requiring models to accomplish two objectives: jointly classify emotions and intentions, and generate explanations for the underlying reasons. Figure 1 illustrates the key distinction between traditional approaches and our proposed JX4MEI task.

A primary obstacle for JX4MEI is the absence of suitable datasets. Existing benchmarks are typically limited in one of two ways (as detailed in Appendix A): they either focus on joint emotion and intent classification without providing explanations (Singh et al., 2022; Liu et al., 2024), or provide explanations for emotion *only*, neglecting the crucial role of intent (Lian et al., 2025; Cheng et al., 2024a; Wang et al., 2025a). To fill this void, we introduce **XMEI-dataset**, the first large-scale benchmark designed for supporting the joint explanation of both emotion and intent in a multimodal context. Comprising 15,461 samples, involving visual, audio, and text modalities, our dataset is constructed using a novel, high-fidelity annotation pipeline. We leverage a Chain-of-Thought strategy (Wei et al., 2022) to guide MLLMs in generating initial rationales, followed by a rigorous human-in-the-loop refinement process. Crucially, experts manually rewrote explanations to remove synthetic artifacts and adjudicated logical inconsistencies, rejecting approximately 15% of candidate samples to ensure quality independent of generator models. This yields a robust dataset justifying the nuanced assignment of 7 emotion and 9 intent categories.

To tackle JX4MEI, we introduce **XMEI-Qwen**, a novel multimodal model built upon the Qwen2.5-7B-Instruct (Yang et al., 2024) backbone. Unlike standard Q-Formers that rely on randomly initialized learnable queries, XMEI-Qwen advances this paradigm by introducing our novel **L**anguage-**Q**uery **Former** (LQ-Former). The core innovation is its use of modality-specific caption embeddings as queries to inject explicit semantic guidance into the feature integration process. By semantically aligning audio and visual features with their corresponding caption, LQ-Former produces enhanced modality representations that are more informative. Empirical results demonstrate that XMEI-Qwen sets a new state-of-the-art on the XMEI-dataset, outperforming competitive baselines in both prediction and explanation generation. Our main contributions can be summarized as:

- We introduce the novel task of **Joint Multimodal Emotion-Intent Explanation and Classification, JX4MEI**. Motivated by the inherent interplay between emotions and intents in human communication, JX4MEI aims to not only identify what users feel and intend, but also explain why, offering a deeper level of interpretability.
- We construct **XMEI-dataset**, the first large-scale benchmark comprising over 15,000 multimodal samples with annotations that provide joint explanations for both emotion and intent.
- We propose **XMEI-Qwen**, equipped with our novel LQ-Former. By replacing generic learnable queries with modality-specific semantic captions, LQ-Former bridges the semantic gap in multimodal feature alignment. Extensive experiments prove that such explicit semantic guidance is critical for complex multimodal reasoning, establishing a new state-of-the-art on the JX4MEI task.

2 Related Work

2.1 Multimodal Emotion and Intent Jointly Understanding

Research in multimodal affective computing has progressed from recognition tasks focusing on *what* is expressed (Lian et al., 2024, 2023a; Zuo et al., 2023; Zhang et al., 2022), to explanation tasks like emotion rationale generation, which addresses *why* an emotion occurred (Lian et al., 2025, 2023b; Cheng et al., 2024a). However, explaining communicative intent remains a largely unexplored area. Furthermore, given that emotion and intent are often intertwined, explaining them in isolation fails to capture the holistic dynamics of user expression. To address this gap, we introduce JX4MEI, a novel task for the joint prediction and explanation of user emotion and intent. To support this, we construct the accompanying XMEI-dataset, designed to enable research into not just *what* users feel and mean, but *why* their expressions manifest in a unified way.

2.2 Multimodal Fusion in MLLMs

Multimodal large language models (MLLMs) have advanced cross-modal reasoning by fusing modalities. Existing alignment strategies generally fall into two paradigms. The first utilizes a simple projection layer to map encoder features directly into

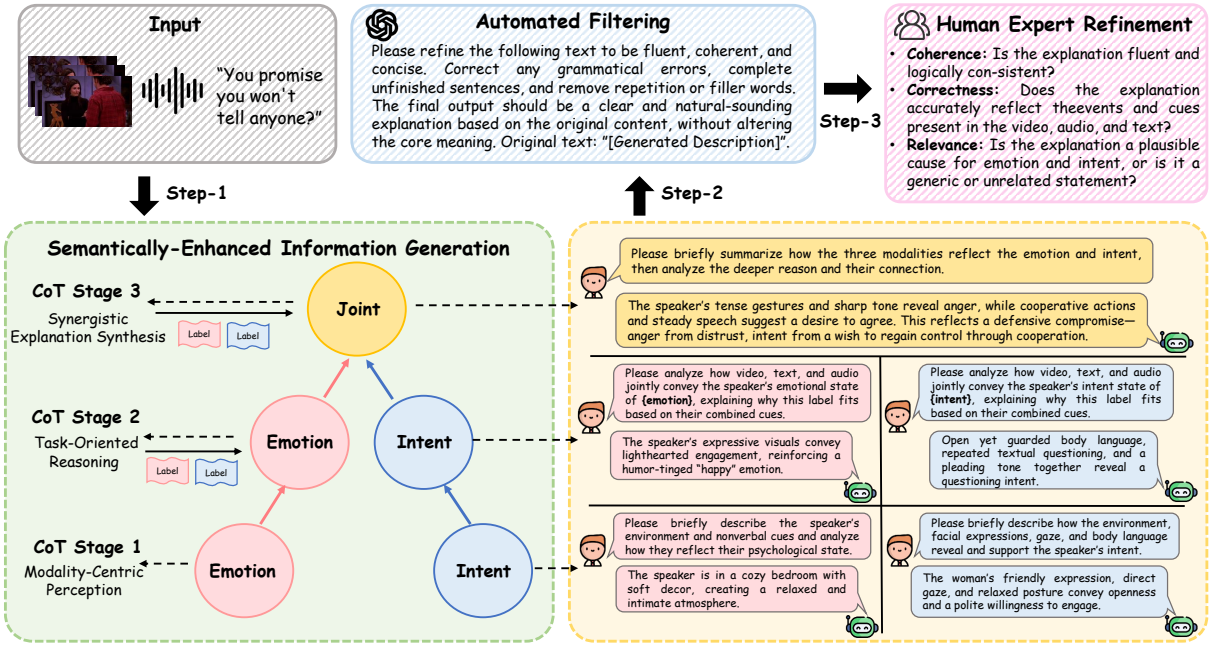


Figure 2: An illustration of our three-stage pipeline for XMEI-dataset construction.

the LLM space, as seen in LLaVA and its variants (Li et al., 2024; Liu et al., 2023). The second employs a Querying Transformer (Q-Former) to distill dense visual features into compact representations via cross-attention (Cheng et al., 2024b), a technique pioneered by BLIP-2 (Li et al., 2023a). However, standard Q-Formers rely on randomly initialized learnable queries, which lack explicit semantics. This limitation hinders their ability to capture fine-grained cues necessary for complex explanation tasks. To address this, we introduce XMEI-Qwen with our novel LQ-Former. This component replaces generic queries with semantic queries constructed from modality-specific descriptions, guiding the model to extract interpretable, task-relevant features and achieving stronger alignment with the reasoning objectives of our proposed JX4MEI task.

3 XMEI-dataset

The JX4MEI task demands not only accurate joint emotion and intent classification but also deep, human-like reasoning. While recent works often rely on direct, one-shot generation from MLLMs (Cheng et al., 2024a; Lian et al., 2023a), such methods frequently fail to produce the structured, step-by-step logic essential for complex causal analysis. To address this, we introduce a specialized **Human-in-the-Loop** Annotation Pipeline to construct our XMEI-dataset. Building upon the MC-EIU dataset (Liu et al., 2024), which provides only classification labels for emotion and intent, we primarily contribute by augmenting a high-quality subset with fine-grained natural language explanations. Specifically, we generate explanations that detail: (1) the distinct causal factors for emotion and intent, (2) the crucial reasoning connecting them. As illustrated in Figure 2, the entire process consists of three stages: Semantically-Enhanced Information Generation, Automated Filtering, and Human Expert Refinement.

Semantically-Enhanced Information Generation. To generate detailed explanations at scale, we employ a three-stage CoT prompting strategy using a suite of specialized models: **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025) for visual understanding, **Qwen2-Audio-7B-Instruct** (Chu et al., 2024) for auditory analysis, and the **QwQ-32B-AWQ** model (Qwen Team, 2025), which we refer to as QwQ, for textual and integrative reasoning. The CoT strategy proceeds as follows:

- **Stage 1: Modality-Centric Perception.** We prompt specialized models to generate objective descriptions for each modality. To prevent data leakage, we enforce strict constraints: models must describe only observable behaviors (e.g., gestures, acoustic features) and are explicitly forbidden from using high-level emotional adjectives or intent-related keywords. This ensures the extracted evidence remains factual and unbiased.

- **Stage 2: Task-Oriented Reasoning.** Using the objective evidence from Stage 1, the QwQ model reasons about the causes of emotion and intent independently. This step prevents premature fusion, ensuring a focused analysis of the distinct causal factors for each state.

- **Stage 3: Synergistic Explanation Synthesis.** The independent rationales from Stage 2 are fed into QwQ to synthesize a final, unified explanation. This stage explicitly articulates the causal interplay, bridging the user's emotion and intent into a coherent narrative.

A complete set of prompts used in this process can be found in Appendix B.

Automated Filtering. To mitigate artifacts in the raw CoT outputs (e.g., redundancy, semantic drift), we employ an automated refinement module using GPT-5 (OpenAI, 2025). This module refines the generated text using a carefully designed prompt (detailed in Appendix C). Crucially, this refinement step is applied sequentially after each of the three CoT generation stages. This prevents error propagation by ensuring that cleaned, coherent output from one stage serves as high-quality input for the next, thereby enhancing the stability and quality of the final explanation.

Human Expert Refinement. Although initialized by MLLMs, our dataset is fundamentally a human-refined augmented dataset. The final and most critical stage involves a meticulous manual verification process conducted by a team of five graduate students with expertise in sentiment analysis and NLP. To ensure the reasoning reflects genuine human cognition rather than model patterns, we moved beyond passive verification to active refinement. First, experts evaluated each explanation against three binary quality criteria: **coherence**, **correctness**, and **relevance**, as detailed in Appendix D. Concurrently, they actively rewrote segments displaying typical LLM artifacts, such as repetitive transitions or overly formal tones, to ensure naturalistic expression. Second, we validated consistency by computing Fleiss' Kappa (Fleiss, 1971), obtaining a score of 0.82, indicating substantial agreement (see Appendix E for full calculation). Moreover, 92.4% of the accepted samples received at least four positive votes out of five across the expert review process, indicating strong consensus beyond the final Fleiss' Kappa score. This provides additional evidence that the retained explanations are not merely plausible, but consistently judged as coherent, correct, and relevant by human annotators. To address the inherent subjectivity in causal reasoning, experts followed a *structured discussion protocol* to resolve disagreements on emotion-intent causality. By prioritizing a *rejection-first policy* for ambiguous or insufficiently grounded links (see Appendix F for detailed procedures and examples), we ensured that the final rationales reflect explicit human consensus rather than majority-vote approximations. This stringent protocol resulted in the removal of approximately 15% of the candidate samples, effectively decoupling the dataset's quality from the generator models' capabilities. Comprehensive statistical analysis of our dataset is available in Appendix G.

4 Method

4.1 Task Definition

Given a multimodal sample (video V, audio A, and text T), JX4MEI requires a model to jointly predict an emotion category C_{emo}, an intent category C_{int}, and generate a joint explanation E for these two mental states. We formulate JX4MEI as a structured generation task where the model learns to produce a single, unified sequence Y that encapsulates all three outputs (the emotion category, intent category, and the explanation), conditioned on a task instruction I_{task}. The ground-truth target Y_{gt} is formatted using a specific template, such as:

> *The user's emotion is {C_{emo}}, intent is {C_{int}}. The underlying reason is: {E}.*

Formally, the model's objective is to generate output $Y = (y_1, y_2, \ldots, y_M)$. The probability of this sequence is modeled autoregressively as:

$$P(Y|V, A, T, I_{\text{task}}) = \prod_{t=1}^{M} P(y_t|y_{<t}, V, A, T, I_{\text{task}}) \tag{1}$$

where y_t is the t-th token, M is the length of the sequence, and I_{task} is a unified instruction prompting the model to jointly identify and explain the speaker's emotion and intent.

4.2 Model Architecture

The architecture of XMEI-Qwen, illustrated in Figure 3, is structured around three key components: (1) a set of frozen multimodal encoders, (2) our novel Language-Query Former (LQ-Former) for

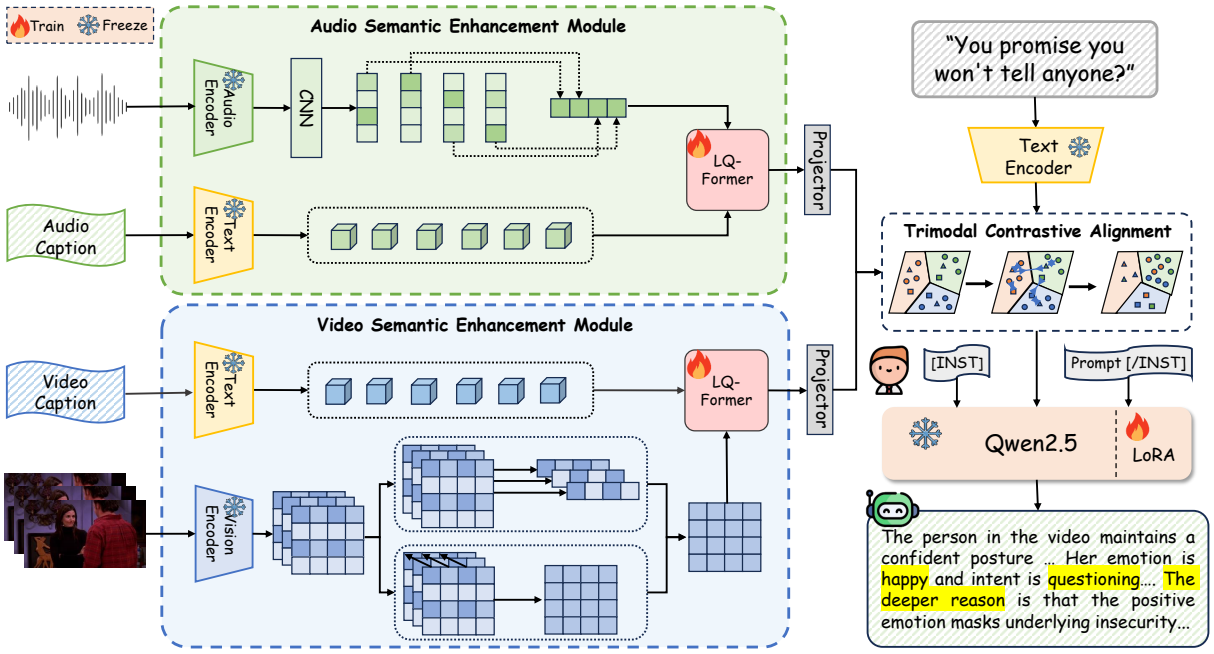


Figure 3: The architecture of our XMEI-Qwen model with novel LQ-Former.

modality enhancement, and (3) a Qwen2.5-7B-Instruct model fine-tuned with LoRA for multimodal fusion and generation. A key distinction of our approach lies in the use of modality-specific captions as auxiliary semantic guidance, which is fused with raw visual and audio features by the LQ-Former to produce enhanced representations.

4.2.1 Multimodal Encoders

We process each modality through a dedicated frozen encoder to extract representations.

Visual Representation. For visual input, we employ a frozen CLIP ViT-L (Radford et al., 2021) to encode N frames of the video, where each frame is decomposed into M patches. Let $p_{i,j} \in \mathbb{R}^d$ be the feature vector for the j-th patch of the i-th frame. To capture comprehensive spatio-temporal dynamics, we extract two complementary feature sets. First, we compute frame-level *spatial features* by mean-pooling all patch features within each frame, yielding a feature vector $\mathbf{f}_{\text{space},i} \in \mathbb{R}^d$ for each frame i:

$$\mathbf{f}_{\text{space},i} = \frac{1}{M} \sum_{j=1}^{M} p_{i,j} \tag{2}$$

These N vectors are then stacked to form the complete spatial feature matrix $\mathbf{F}_{\text{space}} \in \mathbb{R}^{N \times d}$:

$$\mathbf{F}_{\text{space}} = \text{Stack}(\mathbf{f}_{\text{space},1}, \mathbf{f}_{\text{space},2}, \ldots, \mathbf{f}_{\text{space},N}) \tag{3}$$

Second, we derive patch-level *temporal features* by mean-pooling features of the same patch location across all frames. This results in a feature vector $\mathbf{f}_{\text{temp},j} \in \mathbb{R}^d$ for each of the M patch locations:

$$\mathbf{f}_{\text{temp},j} = \frac{1}{N} \sum_{i=1}^{N} p_{i,j} \tag{4}$$

Similarly, these M vectors are stacked to form the temporal feature matrix $\mathbf{F}_{\text{temp}} \in \mathbb{R}^{M \times d}$:

$$\mathbf{F}_{\text{temp}} = \text{Stack}(\mathbf{f}_{\text{temp},1}, \mathbf{f}_{\text{temp},2}, \ldots, \mathbf{f}_{\text{temp},M}) \tag{5}$$

Finally, these two complementary feature matrices are concatenated along their sequence dimension to form the unified visual representation $\mathbf{F}_V \in \mathbb{R}^{(N+M) \times d}$:

$$\mathbf{F}_V = \text{Concat}(\mathbf{F}_{\text{space}}, \mathbf{F}_{\text{temp}}) \tag{6}$$

Audio Representation. A frozen HUBERT-L (Hsu et al., 2021) model and a pooling layer (f_{pool}) encode the audio stream A into a primary feature sequence $\mathbf{F}_A$:

$$\mathbf{F}_A = f_{\text{pool}}\left(\text{HUBERT}(A)\right) \tag{7}$$

4.2.2 LQ-Former for Modality Enhancement

Existing fusion paradigms face limitations in fine-grained reasoning. Standard Q-Formers (Li et al., 2023a) rely on randomly initialized queries, effectively performing *blind learning* that requires

Method	A	V	T	BLEU-4 ↑	ROUGE-L ↑	METEOR ↑	BERTScore ↑	BLEURT↑	CIDEr ↑
SALMONN (Tang et al., 2023)	✓	✗	✓	0.1513	0.4639	0.4730	0.7935	-0.5011	0.4786
PandaGPT (Su et al., 2023)	✓	✗	✓	0.1363	0.4211	0.4187	0.7132	-0.7031	0.4079
Qwen-Audio (Chu et al., 2023)	✓	✗	✓	0.1637	0.4817	0.4791	0.8002	-0.4796	0.4903
Our (A+T)	✓	✗	✓	**0.1712**	**0.4853**	**0.4886**	**0.8156**	**-0.4689**	**0.5021**
Valley (Luo et al., 2023)	✗	✓	✓	0.1437	0.4376	0.4179	0.7010	-0.4072	0.4809
VideoChat (Li et al., 2023b)	✗	✓	✓	0.1501	0.4408	0.4570	0.7137	-0.4339	0.4903
Video-LLaMA (Zhang et al., 2023)	✗	✓	✓	0.1279	0.4096	0.3901	0.7033	-0.6837	0.4479
Video-LLaMA2 (Cheng et al., 2024b)	✗	✓	✓	0.1479	0.4659	0.4083	0.8061	-0.4430	0.5139
Video-LLaVA (Lin et al., 2023)	✗	✓	✓	0.1408	0.4237	0.3971	0.7713	-0.4370	0.4671
Video-ChatGPT (Maaz et al., 2023)	✗	✓	✓	0.1503	0.4479	0.4728	0.7830	-0.5029	0.4660
MiniGPT-4 (Zhu et al., 2023)	✗	✓	✓	0.1370	0.4090	0.4276	0.7513	-0.5371	0.4091
MiniGPT-v2 (Chen et al., 2023)	✗	✓	✓	0.1535	0.4516	0.4159	0.7835	-0.5832	0.4766
Our (V+T)	✗	✓	✓	**0.1624**	**0.4831**	**0.4782**	**0.8115**	**-0.4065**	**0.5284**
PandaGPT (Su et al., 2023)	✓	✓	✓	0.1505	0.4672	0.4711	0.7912	-0.5133	0.4755
Emotion-LLaMA (Cheng et al., 2024a)	✓	✓	✓	0.1970	0.5024	0.5031	0.8278	-0.4098	0.5366
AffectGPT (Lian et al., 2025)	✓	✓	✓	0.2158	0.5193	0.5142	0.8366	-0.3852	0.6021
XMEI-Qwen (Ours)	✓	✓	✓	**0.2415**	**0.5528**	**0.5386**	**0.8692**	**-0.3541**	**0.6284**

Table 1: Main results on the XMEI-dataset test set for explanation generation. We compare our model, **XMEI-Qwen**, against various state-of-the-art multimodal models. Note that AffectGPT utilizes Qwen2.5 but lacks semantic guidance. The best results are in bold and the second best are underlined.

massive data to converge and often misses subtle cues. Conversely, simple projections (e.g., LLaVA (Liu et al., 2023)) risk propagating irrelevant noise from raw features.

To bridge low-level signals with high-level reasoning, we introduce the **Language-Query Former (LQ-Former)**. It is crucial to distinguish LQ-Former from recent text-guided multimodal fusion architectures (e.g., ReWind (Diko et al., 2024)). Architecturally, prior methods typically utilize high-level task instructions to guide a memory module, operating essentially as a *temporal filter* to select which video frames to retain for long-video compression. In contrast, LQ-Former addresses a fundamentally different problem: causal reasoning, by operating explicitly in the *feature space*. Unlike standard approaches, LQ-Former discards random queries in favor of content-aware queries derived from modality-specific captions, as illustrated in Figure 4. These captions, rich in semantic details (e.g., "trembling voice"), act as explicit priors to guide the attention mechanism. Importantly, these semantic queries are restricted to objective, observable cues and explicitly exclude emotion or intent labels, preventing label leakage while still providing informative guidance for feature extraction (see Appendix B.1 for the full prompts). Formally, given a modality feature sequence F_m (e.g., audio or visual) and its caption C_m, we first encode the caption into language-aware queries $Q_L = \text{TextEncoder}(C_m)$. These queries then attend to the modality features via cross-attention:

$$F'_m = \text{CrossAttention}(Q_L, K_m, V_m) \quad (8)$$

where K_m and V_m are projections of F_m. This mechanism performs a targeted distillation, selectively extracting modality features that align with the high-level semantics of the caption. The resulting representation F'_m is thus semantically grounded, significantly facilitating the downstream causal explanation generation.

4.2.3 Multimodal Fusion and Generation

The enhanced visual (F'_V) and audio (F'_A) features from our LQ-Formers, together with the corresponding text features (F_T), are projected into the LLM's embedding space. To improve alignment, we apply a trimodal contrastive loss $\mathcal{L}_{\text{contrastive}}$. The resulting features are then concatenated to form a multimodal soft prompt, which, conditioned on the instruction I_{task}, guides the LLM to generate the structured output Y in an autoregressive manner, as defined in Equation 1. For parameter-efficient fine-tuning, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022). The set of trainable parameters, denoted as θ_{peft}, includes the LoRA adapters, LQ-Formers, and projection layers. The primary training objective is a single, unified generation loss, $\mathcal{L}_{\text{gen}}$, which maximizes the log-likelihood of generating the ground-truth structured output Y_{gt}. By formulating the task this way, the model learns to perform classification by correctly predicting the

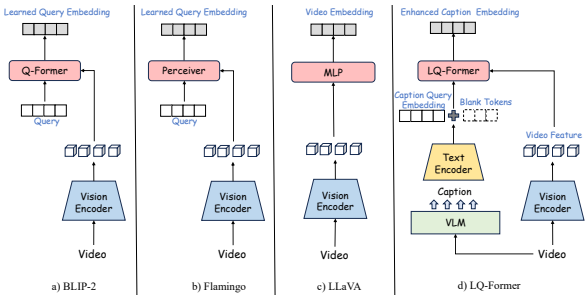


Figure 4: Our LQ-Former (d) contrasted with prior fusion methods. Instead of using learnable queries (a, b) or direct MLP projection (c), our approach uses a semantic caption to guide the extraction of visual features.

emotion and intent tokens, and to generate rationales by reconstructing the explanation tokens, all guided by the same cross-entropy loss objective:

$$\mathcal{L}_{\text{gen}} = -\sum_{i=1}^{N}\sum_{t=1}^{M_i} \log P\big(y_t^{(i)} \mid y_{<t}^{(i)}, V^{(i)}, A^{(i)}, T^{(i)}, I_{\text{task}}^{(i)}; \theta_{\text{peft}}\big) \tag{9}$$

where N is the number of samples and M_i is the length of the i-th target sequence $Y_{\text{gt}}^{(i)}$.

The final training objective is a weighted sum of the generation loss and the contrastive alignment loss, ensuring the model simultaneously learns to align modalities and generate coherent outputs:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda \mathcal{L}_{\text{contrastive}} \tag{10}$$

where λ is a hyperparameter that balances the two objectives. By minimizing $\mathcal{L}_{\text{total}}$, our model learns to produce high-quality outputs that fulfill all requirements of the task.

5 Experiments

5.1 Experimental Setup

Dataset and Baselines. We conduct experiments on the XMEI-dataset, partitioned into speaker-disjoint training (80%), validation (10%), and test (10%) sets. For fair comparison, all baselines (Appendix J) were fully fine-tuned on the training split using their official implementations.

Evaluation Metrics. We evaluate the JX4MEI task on two dimensions: explanation quality and classification accuracy. **(1) Explanation Quality:** We employ a dual strategy for comprehensive assessment. First, we use a suite of automatic metrics, including BLEU-4, ROUGE-L, METEOR, BERTScore, BLEURT, and CIDEr, to measure lexical and semantic similarity with the ground truth. Second, to assess deeper reasoning, we use Gemini 3.0 Pro (Google DeepMind, 2025) as an objective judge to score each explanation on a continuous scale from 0 to 1 on four Task-specific Aspects and four General Quality Aspects (detailed rubrics are provided in Appendix H). **(2) Classification Accuracy:** We adopt the Joint Recognition Balance Metrics (JRBM) from the MEIJU'25 Challenge. JRBM computes the harmonic mean of Micro-F1 scores for emotion and intent, providing a single, balanced score for joint performance (precise formula in Appendix I).

Implementation Details. Our XMEI-Qwen model is built upon the Qwen2.5-7B-Instruct (Yang et al., 2024) backbone. We use CLIP ViT-L (Radford et al., 2021) as the visual encoder and HUBERT-L (Hsu et al., 2021) as the acoustic encoder. The model is fine-tuned using LoRA with a rank (r) of 64. We use a global batch size of 128 and optimize using AdamW with a learning rate of 2×10^{-5} on 4 NVIDIA A800 GPUs.

5.2 Main Results

Explanation Quality Performance. As shown in Table 1, XMEI-Qwen demonstrates comprehensive superiority, consistently outperforming all baselines across standard automatic metrics. Beyond surface-level metrics, the fine-grained LLM evaluation in Table 2 reveals why our model excels. XMEI-Qwen achieves top scores in both general quality aspects and crucial task-specific dimensions. Notably, its high marks in Emotion Clue and Intent Clue confirm its ability to ground explanations in observable multimodal evidence. Furthermore, its leading score in Reason Overlap indicates it accurately captures the causal link between emotion and intent. This dual achievement of fluency and factual grounding confirms that XMEI-Qwen's strong performance is not a result of surface-level mimicry but stems from a genuine improvement in multimodal understanding and logical reasoning.

Joint Labels Prediction Performance. This capacity for high-quality explanation generation is not accidental; it is built upon a foundation of superior predictive accuracy. We extract emotion and intent labels by parsing the generated output. As shown in Table 3, XMEI-Qwen sets a state-of-the-art with a JRBM score of 0.7043, consistently surpassing strong baselines like AffectGPT. Its su-

Method	Task-specific Aspects				General Quality Aspects			
	Label	Emo. Clue	Int. Clue	Reason	Coherence	Fluency	Naturalness	Style
Valley	0.2	0.2	0.2	0.1	0.2	0.5	0.4	0.3
PandaGPT	0.2	0.3	0.3	0.2	0.3	0.6	0.5	0.4
Video-LLaMA2	0.3	0.3	0.4	0.3	0.4	0.7	0.6	0.5
Video-Chat	0.3	0.4	0.4	0.3	0.4	0.7	0.6	0.5
SALMONN	0.3	0.4	0.4	0.3	0.5	0.8	0.7	0.6
MiniGPT-v2	0.4	0.5	0.5	0.4	0.5	0.8	0.7	0.6
Emotion-LLaMA	0.5	0.6	0.5	0.5	0.7	0.8	0.8	0.7
AffectGPT	0.6	0.7	0.6	0.6	0.8	0.9	0.8	0.8
XMEI-Qwen (Ours)	**0.8**	**0.9**	**0.8**	**0.8**	**0.9**	**0.9**	**0.9**	**0.9**

Table 2: LLM-based evaluation of explanation quality, with scores assigned by Gemini 3.0 Pro (0-1 scale). Our model achieves superior scores, particularly in task-specific reasoning dimensions, validating the effectiveness of our semantic-guided approach.

Method	Micro-F1 emotion	Micro-F1 intent	JRBM
Valley	0.2412	0.2367	0.2389
PandaGPT	0.3597	0.3690	0.3643
Video-LLaMA	0.4120	0.4217	0.4169
VideoChat	0.4370	0.4510	0.4439
Video-ChatGPT	0.4660	0.4750	0.4704
SALMONN	0.4937	0.5011	0.4974
Qwen-Audio	0.5512	0.5492	0.5502
MiniGPT-v2	0.5381	0.5677	0.5526
Emotion-LLaMA	0.6301	0.6437	0.6364
AffectGPT	0.6752	0.6814	0.6783
XMEI-Qwen (Ours)	**0.6985**	**0.7102**	**0.7043**

Table 3: Joint emotion and intent classification results on the XMEI-dataset test set. Our model significantly outperforms all baselines.

Model Configuration	LQ-Former	JRBM ↑	CIDEr ↑
Our Full Model			
XMEI-Qwen (Ours)	✓	**0.7043**	**0.6284**
Ablation Variants (Architecture)			
w/o Semantic Enhancement	✗	0.6125	0.5310
w/o Language-Query	✗	0.6580	0.5825
w/o Contrastive Alignment	✓	0.6795	0.6050
Impact of Input Modalities			
w/o Visual Enhancement	Audio Only	0.6690	0.5980
w/o Audio Enhancement	Visual Only	0.6850	0.6095
w/o Text	✓	0.6480	0.5750

Table 4: Ablation study of our XMEI-Qwen model on the XMEI-dataset test set. We evaluate the impact of different architectural modules and input modalities.

perior discriminative ability, attributed to our proposed semantic enhancement architecture, provides the solid foundation necessary for high-fidelity explanation generation.

5.3 Ablation Study

Architectural Components. We first evaluate the contribution of each key component in XMEI-Qwen, with results summarized in Table 4. Removing the entire **Semantic Enhancement Framework** (w/o Semantic Enhancement) causes the most significant performance collapse. In this setting, we bypass our proposed modules entirely; raw features are passed through a simple linear projector. This naive approach results in a dramatic drop of 13.0% in JRBM and 15.5% in CIDEr, confirming that semantic guidance is crucial for this complex reasoning task. Replacing our **LQ-Former** with standard Q-Former (w/o Language-Query) causes drops of 6.6% in JRBM and 7.3% in CIDEr. This demonstrates that our proposed explicit semantic queries are far more effective than blind learning for extracting feature-aligned cues. Visualizations in Appendix K further confirm that LQ-Former effectively grounds semantic concepts (e.g., gestures) while the baseline is distracted by background noise. Furthermore, removing the **Trimodal Contrastive Alignment** (w/o Contrastive Alignment) leads to a notable performance drop of 3.5% in JRBM. This confirms that explicitly pre-aligning multimodal representations in a shared semantic space is critical for effective fusion.

Decoupling Architectural Gains from Information Advantage. To investigate whether XMEI-Qwen's performance stems from its architectural innovation or simply the availability of additional caption information, we conducted a controlled fairness comparison. We augmented the strongest baseline, AffectGPT, by prepending the same modality-specific captions as raw text to its input (AffectGPT + Captions). As shown in Table 5, while

Model	Caption Input	JRBM ↑	CIDEr ↑
AffectGPT (Original)	None	0.6783	0.6021
AffectGPT + Captions	Text-prepend	0.6811	0.6163
XMEI-Qwen (Ours)	**LQ-Former**	**0.7043**	**0.6284**

Table 5: Fairness analysis comparing our LQ-Former against a baseline augmented with the same caption information. The results demonstrate that the architectural design of LQ-Former provides significant gains beyond mere caption availability.

textual captions yield marginal improvements for the baseline (+0.28% JRBM, +1.42% CIDEr), they fail to bridge the gap with XMEI-Qwen, which maintains a substantial lead of +2.32% in JRBM. This suggests that providing captions as LLM context is insufficient for complex reasoning. The strength of LQ-Former lies in its ability to utilize captions as structured semantic queries to perform targeted cross-attention, enabling precise feature-level grounding that simple text-prepending cannot replicate. Furthermore, we evaluate LQ-Former's robustness to caption quality variations by employing a lighter captioning model (BLIP-2). As detailed in Appendix B.4, our model effectively bounds error propagation and consistently outperforms no-caption baselines regardless of the captioning source.

Impact of Input Modalities. To address modality dominance concerns, we conducted input modality ablation studies (see the bottom of Table 4). Removing the Visual modality (w/o Visual Enhancement) results in a JRBM of 0.6690, a drop of approximately 5.0%, confirming the indispensable role of visual cues. Similarly, removing the Audio modality (w/o Audio Enhancement) yields a JRBM of 0.6850, a more modest drop of 2.7%, indicating the unique contribution of acoustic features. Removing the Textual modality (w/o Text) leads to a substantial performance degradation, with a JRBM of 0.6480 (CIDEr: 0.5750), a drop of approximately 8.0% in JRBM and 8.5% in CIDEr. Collectively, these studies confirm our XMEI-Qwen model relies on genuine cross-modal fusion, with each modality contributing uniquely and synergistically to achieve superior emotion and intent understanding.

6 Conclusion

In this work, we address a critical gap in sentiment analysis: the prevalent focus on classification at the expense of understanding the profound interplay between user emotions and intents. We introduce the novel task of **JX4MEI** and the large-scale **XMEI-dataset**, featuring fine-grained, CoT-based annotations. Furthermore, we propose **XMEI-Qwen**, equipped with our semantically-guided **LQ-Former** to enhance multimodal reasoning. Extensive experiments demonstrate that XMEI-Qwen establishes a new state-of-the-art. More importantly, our in-depth ablation studies confirm that this superiority stems from semantically-grounded reasoning rather than surface-level matching. We shift the paradigm from merely identifying *what* emotion and intent are present to explaining *why* they appear together, paving the way for transparent and trustworthy AI systems.

Limitations

While XMEI-Qwen demonstrates superior performance in joint emotion-intent understanding, our current framework operates within the paradigm of discrete categorical classification. Real-world human communication, however, often involves fluid, continuous shifts in affective states that predefined categories may not fully capture. Although our natural language explanations mitigate this by providing nuanced rationales, future work could explore integrating dimensional emotion models (e.g., Valence-Arousal) or open-ended intent descriptions to further enhance the granularity and flexibility of multimodal reasoning.

Ethical Considerations

The XMEI-dataset is constructed by extending the publicly available MC-EIU dataset (Liu et al., 2024). We strictly adhere to the data usage policies and copyright licenses of the original benchmark. As the visual and audio content in MC-EIU originates from movies and TV shows, involving professional actors in fictional scenarios, our dataset avoids the privacy risks typically associated with surveillance footage or real-world user-generated content. We ensure that no Personally Identifiable Information (PII) of non-public individuals is collected or exposed, and the dataset is released solely for academic research purposes.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Nos. 62272092, 62172086), and the Fundamental Research

Funds for the Central Universities under Grants (N25XQD004).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024a. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024b. Video-llama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Anxhelo Diko, Ting Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. 2024. Rewind: Understanding long videos with instructed learnable memory. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13734–13743.

Gunning D Aha DW. 2019. Darpa's explainable artificial intelligence program. *AI Mag*, 40(2):44.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Google DeepMind. 2025. Introducing Gemini 3: A new era of intelligence.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 64–73.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Hui Lee, Singh Suniljit, and Yong Siang Ong. 2025. Dynamic multimodal sentiment analysis: Leveraging cross-modal attention for enabled classification. *arXiv preprint arXiv:2501.08085*.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: What else influences visual instruction tuning beyond data?

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, and 1 others. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, and 1 others. 2023a. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 9610–9614.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, and 1 others. 2023b. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, and 1 others. 2024. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 41–48.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. 2024. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*.

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

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*.

Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 476–489.

M. Minsky. 1986. *The Society of Mind*. Touchstone book. Simon and Schuster.

OpenAI. 2025. Introducing GPT-5.

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

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

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.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300.

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

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Lin Wang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Zhitao Zhang. 2025a. Generative emotion cause explanation in multimodal conversations. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 1394–1403.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2020. Fine-grained emotion and intent learning in movie dialogues. *arXiv preprint arXiv:2012.13624*.

Xingle Xu, Shi Feng, Daling Wang, Yifei Zhang, and Xiaocui Yang. 2025a. Enhancing zero-shot emotion perception in conversation through the internal-to-external chain-of-thought. In *International Conference on Database Systems for Advanced Applications*, pages 209–224. Springer.

Xingle Xu, Yongkang Liu, Dexian Cai, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. 2025b.

Molan: A unified modality-aware noise dynamic editing framework for multimodal sentiment analysis. *arXiv preprint arXiv:2508.09145*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

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*.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1688–1697.

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. *arXiv preprint arXiv:2205.10237*.

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

Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

A Detailed Comparison with Existing Datasets

Table 7 provides a detailed comparison of our XMEI-dataset with several mainstream benchmarks in the field of sentiment analysis. The comparison highlights a critical gap in the field: prior datasets either focus on joint emotion-intent *classification* without explanations (Singh et al., 2022; Liu et al., 2024) or provide explanations for *emotion only*, neglecting intent (Cheng et al., 2024a; Lian et al., 2025). Our XMEI-dataset is the first to address this gap by providing joint explanations for both emotion and intent within a multimodal (Text, Audio, Video) framework. This makes it a unique and essential resource for advancing research on explainable sentiment analysis.

B Prompts for Semantic-Enhanced Information Generation

The following are the prompts used in our three-stage Chain-of-Thought (CoT) strategy. Text enclosed in `<angle brackets>` or `{curly braces}` are placeholders that are programmatically replaced with the relevant model outputs or ground-truth labels during the generation process.

B.1 Stage 1: Modality-Centric Perception Prompts

In the first stage, our primary objective is to extract objective, high-fidelity descriptions of the multimodal inputs to serve as the foundation for subsequent reasoning. Crucially, to prevent **data leakage** and circular reasoning, we enforce strict constraints in the prompts. The specialized models (Qwen2.5-VL for video, Qwen2-Audio for audio, and the QWQ for text analysis) are explicitly instructed to describe only observable physical, acoustic, and linguistic features, and are **strictly prohibited** from using high-level emotional adjectives (e.g., "happy", "angry") or intent-related keywords.

The detailed prompts for each modality, including the specific negative constraints, are presented in Table 6. These objective captions are subsequently used both as the context for the reasoning stages (Stage 2 & 3) and as the semantic queries for the LQ-Former in our XMEI-LLaMA model.

B.2 Stage 2: Task-Oriented Reasoning Prompts

In the second stage, we leverage the reasoning capabilities of the QwQ model to generate the intermediate rationales. We provide the model with the objective descriptions generated in Stage 1 along with the ground-truth classification labels (emotion or intent). The model is prompted to perform **abductive reasoning**: given the observations (multimodal descriptions) and the conclusion (label), it must articulate the logical chain that connects them. We conduct emotion reasoning and intent reasoning separately in this stage to ensure that the model captures the distinct causal factors for each psychological state before synthesizing them.

Emotion Reasoning.

> *Please provide a concise analysis that integrates the following three modalities: Video description:* `<Video`

Modality	Objective Description Prompt
Visual Modality (Input to Qwen2.5-VL)	*"Please provide a purely objective description of the visual content in this video frame. Focus strictly on observable facts. Describe the following elements in detail:* *(1)* ***Environment:*** *The physical setting, lighting conditions, and objects present.* *(2)* ***Appearance:*** *The speaker's clothing and physical characteristics.* *(3)* ***Facial Micro-expressions:*** *Describe specific muscle movements (e.g., corners of mouth raised/lowered, eyebrows furrowed/relaxed, eyes wide/narrow)* ***without*** *naming the emotion.* *(4)* ***Body Language:*** *Describe gestures, posture, and head movements objectively.* ***Constraint:*** *Do not infer the person's mental state or intent. Describe only what is visually visible."*
Audio Modality (Input to Qwen2-Audio)	*"Listen to the audio segment and objectively describe the acoustic characteristics of the speaker's voice. Focus on:* • ***Tone and Pitch:*** *Is the voice high or low pitched? Is the pitch stable or fluctuating?* • ***Volume and Speed:*** *Is the speech loud or soft? Is it fast, slow, or punctuated by pauses?* • ***Vocal Quality:*** *Describe qualities like trembling, breathiness, roughness, or sharpness.* ***Constraint:*** *Avoid using emotional labels (e.g., 'angry tone', 'cheerful voice'). Use descriptive acoustic terms instead."*
Textual Modality (Input to QWQ)	*"Analyze the linguistic structure and semantic content of the speaker's dialogue. Focus strictly on the text itself:* • ***Syntactic Structure:*** *Are the sentences short/long? Are there questions, exclamations, or interruptions?* • ***Word Choice:*** *Note the use of specific types of words (e.g., formal/informal, hesitation markers, strong verbs).* • ***Rhetorical Devices:*** *Identify any visible rhetorical patterns (e.g., repetition, irony) based solely on the text.* ***Constraint:*** *Do not interpret the speaker's underlying intent or emotion yet. Describe only the linguistic patterns observed in the transcript."*

Table 6: Prompts used in Stage 1 for generating modality-specific captions. Note the explicit constraints preventing the generation of emotional labels to ensure zero data leakage.

Datasets	Emotion Label	Intent Label	Modality	Audio Description	Visual Description	Text Description	Emotion Explanation	Intent Explanation	Joint Explanation
MIntRec (Zhang et al., 2022)	✗	✓	T+A+V	✗	✗	✗	✗	✗	✗
Behance-Int (Maharana et al., 2022)	✗	✓	T+A+V	✗	✗	✗	✗	✗	✗
DailyDialog (Li et al., 2017)	✓	✗	T	✗	✗	✗	✗	✗	✗
IEMOCAP (Busso et al., 2008)	✓	✗	T+A+V	✗	✗	✗	✗	✗	✗
MELD (Poria et al., 2018)	✓	✗	T+A+V	✗	✗	✗	✗	✗	✗
M3ED (Zhao et al., 2022)	✓	✗	T+A+V	✗	✗	✗	✗	✗	✗
Twitter (Herzig et al., 2016)	✓	✓	T	✗	✗	✗	✗	✗	✗
ED (Welivita et al., 2020)	✓	✓	T	✗	✗	✗	✗	✗	✗
OSED (Welivita et al., 2020)	✓	✓	T	✗	✗	✗	✗	✗	✗
EmoInt-MD (Singh et al., 2022)	✓	✓	T+A+V	✗	✗	✗	✗	✗	✗
MC-EIU (Liu et al., 2024)	✓	✓	T+A+V	✗	✗	✗	✗	✗	✗
ECEM (Wang et al., 2025a)	✓	✗	T+V	✗	✓	✗	✓	✗	✗
EMER (Lian et al., 2023b)	✓	✗	T+A+V	✓	✓	✓	✗	✗	✗
MERR (Cheng et al., 2024a)	✓	✗	T+A+V	✓	✓	✓	✓	✗	✗
MER-Caption (Lian et al., 2025)	✓	✗	T+A+V	✓	✓	✓	✓	✗	✗
XMEI-dataset(Ours)	✓	✓	T+A+V	✓	✓	✓	✓	✓	✓

Table 7: Detailed comparison of our XMEI-dataset with existing benchmarks in affective computing. This table highlights the unique contributions of our dataset. While prior works typically focus on either emotion/intent classification (e.g., MIntRec, MC-EIU) or provide explanations for emotion only (e.g., MERR, MER-Caption), the XMEI-dataset is the first to provide joint, causal explanations for both emotion and intent in a multimodal (Text, Audio, Visual) setting. This comprehensive annotation scheme, including modality-specific descriptions and dual explanations, makes it the first benchmark specifically designed to support the JX4MEI task.

Description> *Audio description:* <Audio Description> *Text description:* <Text Description> *The ground truth label for the speaker's emotion is* **emotion**. *Based on the above descriptions, infer the deeper connections among the three modalities, analyze why the emotion is classified as* emotion, *and explain how these modalities jointly shape the speaker's emotional state.*

Intent Reasoning.

Please provide a concise analysis that integrates the following three modalities:
Video description: `<Video Description>`
Audio description: `<Audio Description>`
Text description: `<Text Description>`
The ground truth label for the speaker's intent is **`{intent}`**. *Based on the above descriptions, infer the deeper connections among the three modalities, analyze why the intent is classified as* `{intent}`*, and explain how these modalities jointly shape the speaker's communicative intent.*

B.3 Stage 3: Synergistic Explanation Synthesis Prompt

The final stage is designed to synthesize a holistic explanation that bridges the gap between emotion and intent. While Stage 2 analyzes them in isolation, Stage 3 prompts the model to identify the **causal interplay** between them (e.g., how a specific emotion fuels an intent, or how an intent masks an emotion). We feed the separate rationales generated in Stage 2, along with the ground-truth labels, into the QwQ model. The prompt enforces a structured output format to ensure consistency and facilitate downstream training.

Joint Reasoning.

Here are the separate analyses for the speaker's emotion and intent:
Emotion analysis: `<Emotion Rationale from Stage 2>`
Intent analysis: `<Intent Rationale from Stage 2>`
The ground truth emotion label is **`{emotion}`***, and the ground truth intent label is* **`{intent}`**.

Task: *First, synthesize the observations across the three modalities based on the provided analyses. Then, provide a concise but accurate interpretation of the* ***deeper reason*** *why the speaker simultaneously exhibits the emotion* `{emotion}` *and the intent* `{intent}`*. Specifically, explore the* ***causal connection*** *or conflict between the two (e.g., is the emotion driving the intent? Is the intent masking the emotion?).*

Response format:

The person in the video...; the speaker's tone (intonation)...; the speaker's words (text)...

These expressions collectively reflect the speaker's `{emotion}` *emotion and* `{intent}` *intent.*

The deeper reason is: ...

B.4 Robustness Analysis of the Captioning Stage

In this section, we provide a detailed analysis of the model's robustness to caption quality and the mitigation of error propagation, as summarized in Section 5.3 of the main text.

Architectural Constraints on Error Propagation. The captioning stage (Stage 1) is constrained by design to minimize the risk of error propagation. As detailed in the prompts in Table 6 (Appendix B.1), models are strictly prohibited from using emotional adjectives or intent-related keywords, and must describe only objective, observable physical and acoustic behaviors. This architectural separation ensures that even if a captioning model misidentifies a specific gesture or acoustic feature, it cannot directly inject incorrect emotional or intentional labels into the subsequent reasoning pipeline. The errors are thus confined to low-level perceptual descriptions rather than high-level semantic judgments.

Empirical Robustness to Caption Degradation. To quantify the dependency of XMEI-Qwen on high-end captioners (Qwen2.5-VL and Qwen2-Audio), we conducted a degradation experiment using a significantly weaker captioning model, BLIP-2, to simulate a real-world deployment scenario with a lighter pipeline.

As shown in Table 8, even when provided with these lower-quality captions, LQ-Former still yields substantial gains over the no-caption standard Q-Former baseline (+2.41% JRBM and +2.27% CIDEr). While the strongest performance is achieved with our default high-end stack, the controlled performance gap (+2.22% JRBM between weak and strong captioners) demonstrates that LQ-Former is resilient to caption quality variations and provides meaningful semantic guidance regardless of the underlying captioning model.

Caption Source	JRBM ↑	CIDEr ↑
No Captions (Standard Q-Former)	0.6580	0.5825
Weaker Captioner (BLIP-2)	0.6821	0.6052
Ours (Qwen2.5-VL + Qwen2-Audio)	**0.7043**	**0.6284**

Table 8: Robustness analysis of caption quality. LQ-Former demonstrates consistent improvements over the baseline even when driven by a weaker captioning model, confirming its resilience to caption degradation.

C Automated Filtering and Refinement Prompt

During the automated filtering stage, we used GPT-5 with the following prompt to refine the text generated at each step of our CoT pipeline. The placeholder [Generated Description] was replaced with the actual text from the model.

> *Please refine the following description to ensure it is fluent, coherent, and free of repetition. Correct any grammatical errors and complete any unfinished sentences. The final text should be a clear and concise explanation based on the original content: "[Original Description]".*

D Human Expert Verification Criteria

During the manual verification process, our expert annotators evaluated each sample based on the following three criteria. A sample had to pass all three to be considered for inclusion in the final dataset.

- **Coherence:** The explanation must be grammatically correct, fluent, and logically consistent. The flow of reasoning should be easy to follow.
- **Correctness:** The explanation must accurately reflect the events, objects, speech, and non-verbal cues present in the visual, audio, and textual modalities of the sample. It should not fabricate or misrepresent information.
- **Relevance:** The reasoning provided in the explanation must be a plausible cause for the specific emotion-intent pair identified in the sample. It should not be a generic statement or an observation unrelated to the emotional and intentional state of the speaker.

E Inter-Expert Agreement Analysis

To ensure the reliability of our human verification process, we conducted a rigorous inter-annotator agreement analysis. As described in Section 3, five expert annotators independently evaluated each explanation on three binary criteria: *Coherence*, *Correctness*, and *Relevance*.

Agreement Statistics. Table 9 details the distribution of voting patterns across the 15,461 accepted samples. A sample is considered "Accepted" only if it receives at least 3 positive votes ("Yes") out of 5 for **all three criteria**. The high proportion of unanimous (5-Yes) and strong majority (4-Yes) votes demonstrates the high quality of the filtered dataset. Specifically, 92.4% of the accepted samples received at least 4 positive votes, indicating strong consensus among annotators.

Vote Distribution (Yes / No)	Count	Percentage
Unanimous Agreement (5 / 0)	8,967	58.0%
Strong Majority (4 / 1)	5,319	34.4%
Simple Majority (3 / 2)	1,175	7.6%
Total Accepted Samples	**15,461**	**100%**

Table 9: Distribution of annotator votes for the samples included in the final XMEI-dataset. The statistics reflect the consensus level on the overall quality (passing all three criteria).

Fleiss' Kappa Calculation. We calculated Fleiss' Kappa (κ) to measure the reliability of agreement beyond chance. Since each sample was evaluated on three distinct criteria, we first computed the κ score for each criterion independently and then averaged them to obtain the final reported score of 0.82. The breakdown is as follows:

- **Coherence:** $\kappa = 0.85$ (Almost Perfect Agreement)
- **Correctness:** $\kappa = 0.81$ (Substantial Agreement)
- **Relevance:** $\kappa = 0.79$ (Substantial Agreement)

The overall Fleiss' Kappa is the arithmetic mean: $\kappa_{overall} = (0.85 + 0.81 + 0.79)/3 \approx 0.82$. This confirms that the high agreement is consistent across all dimensions of evaluation.

F Disagreement Resolution Protocol

To ensure the reliability of subjective causal links between emotion and intent, our experts followed a

structured protocol to handle cases where the initial agreement was low (fewer than 4 out of 5 "Yes" votes on the *Relevance* criterion).

Resolution Procedure. Disagreements were categorized into two types:

1. **Causal Direction Ambiguity:** Cases where the causal relationship (e.g., does the emotion fuel the intent, or does the intent mask the emotion?) was unclear. These were resolved by collectively re-evaluating specific multimodal cues, such as the micro-expressions' timing relative to speech.

2. **Grounding Insufficiency:** Cases where the causal link was plausible but lacked unique evidence in the video/audio.

Rejection-First Policy. Following a strict consensus model, any sample falling into Category (2), or Category (1) where consensus could not be reached, was rejected rather than adjudicated. This ensures the dataset contains only high-confidence causal rationales.

Representative Examples.

- **Resolved Case:** Annotators initially disagreed on whether a speaker's "tense posture" (anxiety) was the cause of "rapid questioning" (intent: seeking reassurance). Discussion of the audio pitch contour confirmed the anxiety preceded and drove the questioning, leading to a "Resolved" status.

- **Rejected Case:** A speaker showed "disgust" and "refusing." While the explanation linked them logically, the experts found the visual cues for disgust too subtle to uniquely justify it as the sole driver of the refusal. The sample was rejected to maintain the gold-standard quality.

G Detailed Dataset Statistics

This section provides a detailed statistical overview of the final XMEI-dataset, focusing on the 12,369 verified training samples, which constitute 80% of the total dataset.

Label Distribution. Our dataset is annotated with 7 primary emotion categories (happy, surprise, sad, disgust, anger, fear, and neutral) and 9

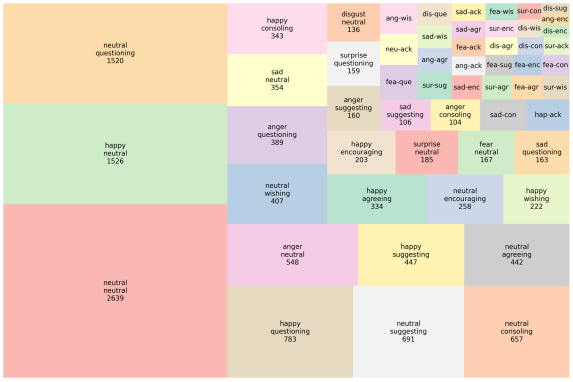


Figure 5: Joint distribution of emotion-intent pairs in the XMEI-dataset, visualized as a treemap.

communicative intents (questioning, agreeing, acknowledging, sympathizing, encouraging, consoling, suggesting, wishing, and neutral). Figure 5 illustrates the joint distribution of these labels, which exhibits a naturalistic, long-tail pattern, characteristic of real-world human communication where a few common states dominate, followed by a wide array of more nuanced interactions. Specifically, the neutral-neutral (2639 samples), happy-neutral (1526), and neutral-questioning (1520) pairs constitute the most frequent categories, reflecting common conversational scenarios. Crucially, the dataset's value lies not only in these high-frequency pairs but also in its extensive coverage of a long tail of less frequent but complex combinations (e.g., anger-consoling, disgust-suggesting, and fear-wishing). This diverse and imbalanced distribution makes the XMEI-dataset a realistic and challenging benchmark, compelling models to perform well on both common-sense scenarios and rare, subtle social cues.

Explanation Statistics. To quantify the depth and complexity of our annotations, we performed a detailed textual analysis. For each sample, our dataset provides three distinct types of explanations: (1) an explanation for the emotion, (2) an explanation for the intent, and (3) a final, synergistic explanation that holistically connects the two. Table 10 presents a comparative statistical analysis of these components, revealing a key feature of our annotation design. While the individual rationales for emotion and intent are themselves substantial (averaging over 50 words each), the synergistic explanation demonstrates a clear leap in complexity, evident across three key metrics:

- **Substantial Length and Added Content:**

Statistic	Emotion Rationale	Intent Rationale	Synergistic Explanation
Avg. Length (words)	45.3	58.1	78.5
Avg. Length (sents)	1.8	2.1	4.2
Lexical Density (%)	45.2	46.1	48.5

Table 10: Textual statistics of the generated explanations in the XMEI-dataset.

The synergistic explanation's average word count of 78.5 is significantly greater than the length of the individual rational (averaging 50). This indicates that nearly 23% of its content consists of new, integrative reasoning rather than simple repetition.

- **Complex Grammatical Structure:** This added content is structured into more complex sentences. With an average of 4.2 sentences, the synergistic explanation moves beyond simple statements to build a narrative, establish relational links, and articulate the nuanced, step-by-step logic connecting emotion and intent.

- **Higher Information Density:** Crucially, this additional text is of high quality. The synergistic explanation achieves the highest Lexical Density (48.5%), demonstrating that the language used to bridge the two concepts is not mere filler but is more precise, content-rich, and uses a more sophisticated vocabulary to convey complex relationships.

This resulting "surplus" in length, structure, and density is not redundant; it represents the additional, complex logic required to bridge the two concepts and articulate their interplay. Collectively, these statistics confirm that our dataset provides a rich and challenging benchmark for training models capable of deep, multi-faceted explanatory reasoning.

H Detailed LLM-based Evaluation Criteria

Automatic metrics may not fully capture the logical coherence and multimodal grounding required by the MEIE task. To address this, we introduce a fine-grained, LLM-based evaluation protocol. We divide our evaluation into four task-specific aspects and four general quality aspects:

Task-specific Aspects

- **Label overlap:** Whether the explanation correctly implies the ground-truth emotion and intent labels.

- **Emotion clue overlap:** Whether the explanation correctly references the key multimodal evidence (e.g., "smiling face") for the emotion.

- **Intent clue overlap:** Whether the explanation correctly references the evidence (e.g., "encouraging words") for the intent.

- **Reason overlap:** Whether the causal reasoning connecting the emotion, intent, and evidence is logical and consistent with the ground truth.

General Quality Aspects

- **Coherence:** Whether the explanation is internally consistent and presents a clear, logical flow of argument, independent of the reference.

- **Fluency:** Whether the explanation is grammatically correct and easy to understand.

- **Naturalness:** Whether the explanation reads like it was written by a human, avoiding robotic, repetitive, or overly formulaic phrasing.

- **Style:** Evaluates the appropriateness of the tone and phrasing for the given context.

I Joint Recognition Balance Metric (JRBM)

To evaluate the model's ability to jointly recognize emotion and intent, we adopt the Joint Recognition Balance Metric (JRBM) from the MEIJU'25 Challenge@ICASSP 2025. It is defined as the harmonic mean of the Micro F1-scores for the two subtasks, ensuring a balanced measure of performance. The formula is given by:

$$JRBM = \frac{2 \times M_{\text{emotion}} \times M_{\text{intent}}}{M_{\text{emotion}} + M_{\text{intent}}} \quad (11)$$

where M_{emotion} and M_{intent} represent the Micro F1-score for the emotion and intent recognition subtasks, respectively.

J Baseline Models

To comprehensively evaluate the performance of our proposed XMEI-Qwen model, we compare it against a range of state-of-the-art multimodal large language models (MLLMs). Detailed descriptions of the models are provided below.

Audio-Text Baselines. For the audio-text modality setting, we compare against leading models designed for speech and audio understanding:

- **SALMONN** (Tang et al., 2023): An MLLM capable of understanding and responding to general audio inputs, including speech, music, and sound events.
- **PandaGPT** (Su et al., 2023): A model that aligns different modalities, such as video and audio, with LLMs, enabling instruction-following capabilities across them.
- **Qwen-Audio** (Chu et al., 2023): A large audio-language model that accepts audio in various forms and performs multiturn dialogue based on it.

Video-Text Baselines. For the video-text setting, which is the most common in multimodal research, we select a diverse set of strong baselines:

- **Valley** (Luo et al., 2023): An MLLM that integrates visual and linguistic features for enhanced video understanding and conversation.
- **VideoChat** (Li et al., 2023b): A video-centric conversational model designed for in-depth understanding of video content.
- **Video-LLaMA** and **Video-LLaMA2** (Zhang et al., 2023; Cheng et al., 2024b): A series of models that integrate a video branch with an LLM to enable video-grounded dialogue.
- **Video-LLaVA** (Lin et al., 2023): A model that extends the LLaVA architecture to the video domain for instruction-tuned visual understanding.
- **Video-ChatGPT** (Maaz et al., 2023): A video conversation model capable of generating detailed descriptions and answering questions about videos.
- **MiniGPT-4** and **MiniGPT-v2** (Zhu et al., 2023; Chen et al., 2023): A series of lightweight yet powerful MLLMs known for their strong visual comprehension and conversational abilities.

Audio-Video-Text Baselines. For the full tri-modal setting, we compare against models capable of processing all three modalities. Note that some models, like PandaGPT, are versatile and appear in multiple settings.

- **PandaGPT** (Su et al., 2023): As mentioned above, this model can handle audio, video, and text inputs for instruction following.
- **Emotion-LLaMA** (Cheng et al., 2024a): A strong baseline specifically designed for multimodal emotion recognition and explanation, making it a highly relevant and competitive model for our task.

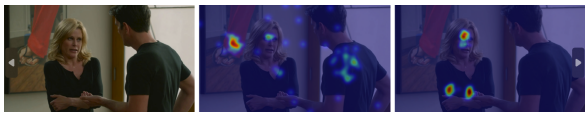
K Detailed Visualization of Attention Mechanisms

In this section, we provide a detailed visual analysis to demonstrate the effectiveness of our proposed Language-Query Former (LQ-Former). Figure 6 presents a side-by-side comparison of the cross-attention maps generated by the Standard Q-Former (Baseline) and our LQ-Former across four distinct samples from the XMEI-dataset.

The visualization reveals a critical limitation of standard learnable queries: they tend to overfit to visually salient objects (Salient Object Bias). For example, in Case 4 (bottom-right), the baseline attention is scattered across the colorful kitchenware and the fridge, missing the speaker's expression entirely. In contrast, by leveraging semantic captions (e.g., *"leaning forward"*, *"surprised face"*) as queries, LQ-Former demonstrates robust **semantic grounding**. It consistently focuses on the regions described in the text—such as the eyes, mouth, and hand gestures—effectively filtering out complex background distractions. This confirms that our explicit semantic guidance strategy enables the model to extract more task-relevant features for emotion and intent understanding.

Appendix B.2: Sample Entry from XMEI-dataset	
Video Frames	
Subtitle	"AND YOU DECIDE THAT THAT'S THE PERFECT MOMENT"
Emotion Label	**Happy**
Intent Label	**Suggesting**
Video Caption	The speaker gestures actively with his hand, displaying raised eyebrows and a wide-open mouth while speaking, maintaining direct eye contact.
Audio Caption	Fast speech rate with high volume. The pitch contour shows a continuous upward inflection, with distinct stress on emphatic words.
Emotion Reason	The visual cues of active gesturing and wide eyes, combined with the fast speech rate and loud volume, indicate a state of high arousal. The text content "perfect moment" carries a positive semantic valence. Therefore, the high energy likely stems from excitement or satisfaction rather than anger.
Intent Reason	The speaker is forcefully describing a specific scenario. The emphatic tone and the gesture of presenting an idea suggest he is not just stating a fact but proposing a perspective to convince the listener, effectively suggesting a shared viewpoint.
Joint Reason	The speaker's emotion is **Happy**, and intent is **Suggesting**.**The deep reason is:** The speaker derives internal satisfaction from identifying what he believes is the "perfect moment," leading to a happy, high-energy state. He channels this excitement into his intent to suggest. By using an enthusiastic tone and emphatic gestures, he attempts to frame his suggestion not as a demand, but as a shared, exciting opportunity, thereby encouraging the listener to agree with his timing.

Table 11: A sample from the XMEI-dataset. The dataset provides separate granular explanations for emotion and intent, along with concise, objective modality captions used for training the LQ-Former, and a final synergistic explanation connecting the two states.



Input *Q-Former* **Ours**

Caption: A slow speaking rate accompanies her defensive stance, marked by furrowed eyebrows and crossed arms

Input *Q-Former* **Ours**

Caption: He speaks with a high pitch through wide eyes, an open mouth, and a sharp pointing finger

Input *Q-Former* **Ours**

Caption: A low, steady tone coincides with the speaker leaning back, featuring narrowed eyes and holding a glass.

Input *Q-Former* **Ours**

Caption: A fast tempo synchronized with an open mouth and the body leaning forward towards the listener

Figure 6: **Visualizing Semantic Guidance.** Qualitative comparison of feature alignment across diverse multimodal scenarios. Each panel displays the **Input Frame** (left), **Standard Q-Former Attention** (middle), and **Our LQ-Former Attention** (right). The caption below is the semantic query, strictly limited to observable acoustic (e.g., pitch, tempo) and visual (e.g., gestures, expressions) features to avoid label leakage. While the standard Q-Former is often distracted by background clutter (e.g., the red shirt in Case 2, colorful knives in Case 4), LQ-Former precisely grounds these physical descriptions into corresponding visual regions.

Appendix C: Qualitative Comparison with Baseline Models

Subtitle	"So...not an edible arrangement."
Ground Truth	The user's emotion is **Happy**, and intent is **Consoling**. **Reasoning:** The speaker is in a relaxed home environment wearing a humorous costume. While the text implies a critique ("not an edible"), her tone is steady but gentle. She uses this humorous metaphor to soften the negative feedback, transforming potential frustration into empathy to comfort the listener.
Valley	The video shows a person speaking. The tone is calm. The emotion seems **neutral**. The speaker is providing information about an arrangement.
SALMONN	It is difficult to predict the emotional state of the individual without more context. However, based on the audio, the tone is steady and calm. It is possible that the individual is feeling **neutral** or **serious**. Alternatively, based on the text "not", they may be feeling **dissatisfied**. Without visual confirmation, it is impossible to accurately predict the intent.
Video-LLaMA	Based on the subtitles, the speaker is saying that something is "not an edible arrangement," which implies a failure or a mistake. The speaker appears to be **disappointed** or **frustrated** because the result is not what was expected. The intent is **complaining** or **criticizing** the outcome.
Video-ChatGPT	In the video, a woman is seen wearing a costume resembling a hot dog or food item. The environment appears to be an indoor setting. The acoustic information is consistent with speech. The subtitles indicate she is making a statement about an arrangement. Based on these cues, the emotion is likely **neutral**. The intent is simply **stating a fact** or **informing** the listener.
PandaGPT	The person in the video is wearing a funny costume and seems to be in a good mood. The emotion is **Happy** or **Amused**. The intent is likely **joking** or **entertaining** the audience because of the silly outfit.
Emotion-LLaMA	In the video, the person's facial expressions and costume primarily suggest a **Happy** emotional state. However, a detailed analysis of her verbal content reveals a critique ("not edible"). These conflicting signals suggest a mixed emotional state. Therefore, it is inferred that she is using **Sarcasm** or **Mocking** to make fun of the mistake, rather than being genuinely happy.
XMEI-Qwen (Ours)	The user's emotion is **Happy**, and intent is **Consoling**. **The deeper reason is:** The apparent humor (costume) is not just for entertainment, but a **strategic social tool** to create a lighthearted atmosphere. This allows the speaker to deliver a critique ("not edible") without causing distress. The steady, gentle tone works with the visual humor to soften the blow, transforming potential frustration into empathy to **comfort** the listener.

Table 12: A qualitative comparison on a challenging sample requiring irony detection. Baseline models exhibit distinct failure modes: **Valley** and **Video-ChatGPT** fail to capture emotional nuance; **Video-LLaMA** relies heavily on negative text; **PandaGPT** focuses superficially on visual humor; **Emotion-LLaMA** misinterprets the multimodal conflict as sarcasm/mocking. Only **XMEI-Qwen** correctly identifies the strategic use of humor to **console**.