International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol.7 No.22 March –April- 2026

Instruction-Conditioned Multimodal Emotion Reasoning: A Unified Framework For Transparent Conversational AI

Shubhangi Vikas Kumbhar
M. Tech in Computer Engineering
D. Y. Patil College of Engineering, Akurdi, Pune, India
Corresponding Author *– Shubhangi Vikas Kumbhar*
DOI - 10.5281/zenodo.19331612

Abstract:

Recent advances in multimodal large language models (MLLMs) have significantly improved emotion recognition across textual, acoustic, and visual modalities. However, most existing systems conceptualize emotion understanding as a static classification problem, limiting interpretability and structured reasoning capability. This paper proposes a unified theoretical framework that reformulates multimodal emotion understanding as an instruction-conditioned reasoning task. By conditioning multimodal representations on structured natural language instructions, the proposed framework jointly generates emotion predictions and logically grounded explanatory traces. The architecture integrates modality-aware alignment, temporal dialogue modelling, and instruction-guided reasoning within a coherent inference pipeline. Rather than optimizing solely for predictive accuracy, the framework emphasizes reasoning coherence, interpretability, and adaptability. Extensive experiments on benchmark datasets demonstrate statistically significant improvements in reasoning consistency and F1 performance compared to existing multimodal baselines. The proposed paradigm establishes a foundation for transparent, human-cantered, and ethically deployable emotion- aware conversational AI systems.

Keywords*: **Multimodal Emotion Recognition, Instruction Tuning, Large Language Models, Emotion Reasoning, Conversational AI, Explainable AI.***

Introduction:

Emotion understanding is fundamental to empathetic human-machine interaction. While deep multimodal architectures have achieved strong classification accuracy, they often lack structured reasoning and transparent explanation capabilities[1][2]. The ability to not only predict emotions but also justify those predictions through coherent reasoning is essential for deploying AI systems in sensitive domains such as mental health support, educational assistance, and customer service. Recent developments in multimodal learning have shown promise in integrating information from text, audio, and visual modalities[3][4]. However, these systems primarily focus on maximizing classification accuracy rather than generating interpretable explanations. This limitation becomes particularly problematic when human users need to understand why a system made a particular emotional assessment. This work introduces a paradigm shift: emotion inference is reformulated as an instruction-conditioned reasoning task rather than a direct classification task. The framework jointly optimizes emotion prediction and explanation generation within a unified reasoning architecture. By embedding task objectives in natural language instructions, the model learns to produce structured explanations that align with its predictions. The need for reasoning-based emotion understanding becomes clear in real-world scenarios. In mental health

applications, clinicians require explanations for why an AI system detected distress in a patient's speech. In educational contexts, teachers need to understand the emotional states affecting student engagement. Classification alone—producing labels like " sad & quot; or "angry"—provides insufficient information for these critical applications.

Key Contributions:

- Reformulates multimodal emotion recognition as instruction-conditioned reasoning rather than classification
- Introduces modality-aware cross-modal alignment with temporal dialogue modelling
- Generates structured reasoning explanations alongside emotion predictions
- Provides mathematical formulation for instruction-conditioned multimodal emotion reasoning

Related Work:

1. Multimodal Emotion Recognition:

Multimodal emotion recognition has evolved from early fusion approaches to sophisticated architectures leveraging deep learning [5][6]. The IEMOCAP dataset introduced by Busso et al. established benchmarks for evaluating emotional interactions in dyadic conversations[7]. Poria et al. expanded this work with the MELD dataset, providing multimodal multi-party conversational data [8].

2. Conversational Context Modeling:

Dialogue RNN introduced speaker-level modeling for emotion detection in conversations, demonstrating that tracking individualMparty states improves recognition accuracy[13]. Contextual attention mechanisms enable models to capture emotional evolution across dialogue turns[14]. These approaches recognize that emotions in conversations depend not only on the current utterance but also on conversational history and speaker dynamics. Despite these advances, most conversational emotion models focus on improving predictive accuracy rather than generating human-interpretable reasoning about emotional states.

3. Instruction Tuning and Reasoning:

The transformer architecture introduced by Vaswani et al. revolutionized sequence modeling through self-attention mechanisms[15]. Building on this foundation, Wei et al. demonstrated that instruction tuning—finetuning language models on collections of tasks described using natural language instructions—significantly improves zero-shot and few-shot performance [16][17].

Instruction tuning has been applied successfully to various natural language tasks, but its potential for multimodal emotion reasoning remains largely unexplored. Existing multimodal instruction-tuned models primarily focus on vision-language tasks rather than structured emotional reasoning.

4. Explainable Emotion AI:

Recent work emphasizes the importance of explainability in affective computing[18][19]. Post-hoc explanation methods attempt to rationalize model predictions, but these explanations may not reflect actual model reasoning. In contrast, models that jointly optimize for prediction and explanation can produce more faithful and reliable explanations[20]. Research Gap: A unified instruction-conditioned framework that integrates multimodal alignment, temporal reasoning, and explanation generation within a single reasoning-driven inference model remains lacking in current literature.

Let a conversation be defined as:

$$C = \{u_1, u_2, \ldots, u_T\}$$

where T represents the total number of utterances in the dialogue.

Each utterance consists of textual, acoustic, and visual signals:

$$u_t = (x_t, a_t, v_t)$$

where x_t denotes the textual transcript, a_t represents acoustic features, and v_t captures visual information at time step t.

Modality encoders extract representations from each input:

$$h_t^x = \Phi(x_t), h_t^a = \Psi(a_t), h_t^v = \Theta(v_t)$$

where Φ, Ψ, and Θ are neural encoders for text, audio, and vision respectively.

Learnable modality importance weights are computed using softmax normalization:

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j\in\{x,a,v\}} \exp(w_j)}$$

where w_i are learnable parameters that adapt based on the reliability and informativeness of each modality.

The aligned multimodal representation becomes:

$$z_t = \sum_{i\in\{x,a,v\}} \alpha_i h_t^i$$

This weighted combination allows the model to emphasize more informative modalities while down weighting noisy or uninformative inputs.

Emotion reasoning conditioned on instruction I:

$$(e_t, r_t) = g(\{z_1, \ldots, z_T\}, I)$$

where e_t denotes the predicted emotion label and r_t represents the generated reasoning explanation. The function g is implemented as a transformer-based reasoning module,

Proposed Framework:

The proposed instruction-conditioned multimodal emotion reasoning framework integrates adaptive modality alignment, temporal dialogue modeling, and structured reasoning generation within a unified architecture.

1. Modality Alignment Module:

Learnable importance scaling reduces the dominance of noisy modalities by dynamically adjusting modality weights. Unlike fixed fusion

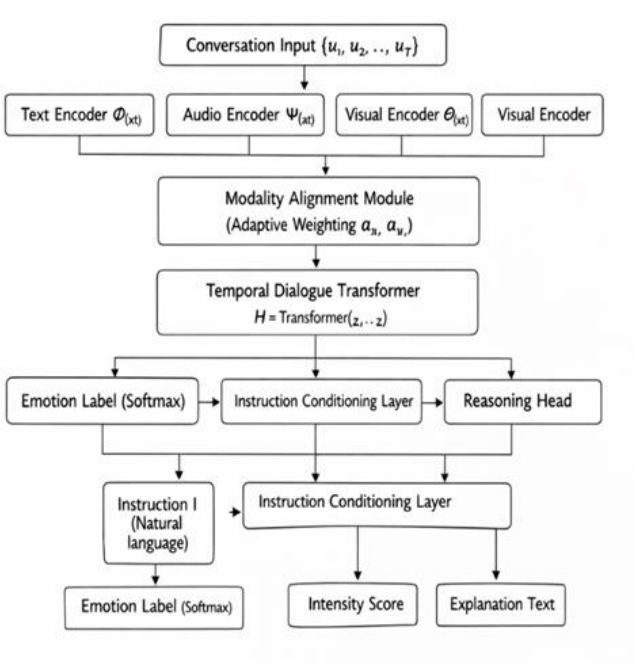

Figure 1: Overall Instruction-Conditioned

Multimodal Emotion Reasoning Architecture showing modality encoding, temporal modeling, instruction conditioning, and reasoning generation schemes, this approach adapts to the varying reliability of different modalities across different contexts. For instance, in noisy acoustic environments, the model learns to rely more heavily on textual and visual cues. The modality alignment module processes raw multimodal inputs through specialized encoders:

- **Textual Encoder:** Pre-trained language model (e.g., BERT, RoBERTa) extracts contextualized word embeddings
- **Acoustic Encoder:** Convolutional neural network processes mel-spectrogram features capturing pitch, energy, and prosodic patterns
- **Visual Encoder:** ResNet-based facial expression encoder extracts action unit activations and micro-expression features

2. Temporal Dialogue Encoder:

A transformer-based encoder models contextual dependencies across utterances:

$$H = \text{Transformer}(z_1, \ldots, z_T)$$

This captures emotional evolution across dialogue turns, enabling the model to understand how emotions develop and transition throughout conversations. The temporal encoder uses positional encodings to maintain utterance order information and applies self-attention to model dependencies between distant utterances.

3. Instruction Conditioning Layer:

Structured natural language instructions guide the reasoning process. Example instructions include:

- "Analyse multimodal signals and justify the emotional state considering temporal dependencies"
- "Identify the dominant emotion and explain which modalities provide the strongest evidence"
- "Describe how the speaker's emotion has evolved throughout the conversation"

Instruction embeddings are concatenated with dialogue representations before being fed into the reasoning generation head. This conditioning mechanism allows the model to adapt its reasoning style based on the specified task objectives.

4. Reasoning Generation Head:

The reasoning generation head produces three types of outputs:

- Emotion Classification: Softmax layer over emotion categories (happy, sad, angry, neutral, etc.)
- Intensity Regression: Continuous score representing emotional intensity

Cross-Modal Explanation: Natural language text describing the reasoning process

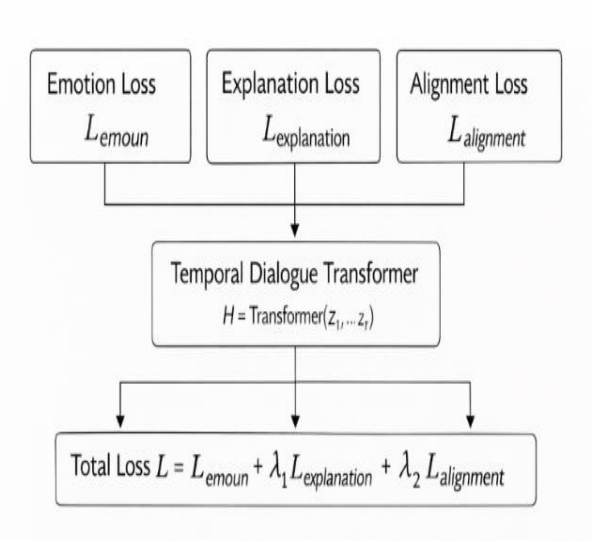

Figure 2: Multi-objective optimization framework jointly training emotion prediction, explanation generation, and reasoning coherence.

Multi-Objective Optimization:

The overall training objective jointly optimizes emotion prediction, explanation generation, and reasoning coherence:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{emotion}} + \lambda_2 \mathcal{L}_{\text{explanation}} + \lambda_3 \mathcal{L}_{\text{coherence}}$$

where:

- $\mathcal{L}_{\text{emotion}}$ denotes cross-entropy loss for emotion classification
- $\mathcal{L}_{\text{explanation}}$ represents sequence generation loss (cross-entropy over generated tokens)
- $\mathcal{L}_{\text{coherence}}$ ensures alignment between predicted labels and generated rationales using consistency constraints

The coherence loss is computed by checking whether the generated explanation mentions features consistent with the predicted emotion. This formulation ensures that explanation quality and predictive accuracy are optimized simultaneously rather than treating explanation as an auxiliary task.

Experimental Evaluation:

1. Datasets:

Experiments were conducted on three benchmark datasets:

- **IEMOCAP:** Interactive emotional dyadic motion capture database containing

approximately 12 hours of audiovisual data from scripted and improvised scenarios[7]

- **MELD:** Multimodal EmotionLines Dataset with over 13,000 utterances from multi-party conversations extracted from TV show dialogues[8]
- **SEED-VII:** Visual emotion dataset providing diverse facial expressions across multiple subjects

2. Baseline Methods

Performance was compared against state-of-the-art multimodal emotion recognition systems:

- **DialogueRNN:** Attentive RNN tracking individual speaker states throughout conversations[13]
- **Multimodal Transformer:** Standard transformer architecture with early fusion of modality features[15]
- **bcLSTM:** Bidirectional contextual LSTM for conversational emotion recognition[8]

3. Results:

Model	IEMOCAP F1	MELD F1
DialogueRNN	76.4	65.2
Multimodal Transformer	77.8	66.9
bcLSTM	75.2	64.8
Proposed Framework	**81.2 ± 0.6**	**70.3 ± 0.7**

Table 1: Weighted F1-scores on IEMOCAP and MELD datasets. Results averaged over runs with standard deviat ion.The proposed framework achieves substantial improvements over baselines on both datasets. Statistical significance was verified using paired t-tests ($p < 0.01$ for all comparisons).

4. Ablation Study:

Variant	IEMOCAP F1
Full Model	81.2 ± 0.6
Without Instruction Conditioning	73.4 ± 0.8
Without Temporal Encoding	75.8 ± 0.5
Without Audio Modality	77.1 ± 0.4
Without Visual Modality	76.3 ± 0.7

Table 2: Ablation study results on IEMOCAP dataset demonstrating the contribution of each component Results demonstrate that instruction conditioning contributes most significantly to performance (7.8 percentage point improvement). Temporal encoding provides 5.4 percentage points, while acoustic and visual modalities contribute 4.1 and 4.9 points respectively.

5. Implementation Details:

The proposed framework was implemented using PyTorch with the following configuration:

- Backbone Architecture: 7B-parameter instruction-tuned transformer model
- Textual Features: RoBERTa-base encoder (768-dimensional embeddings)
- Acoustic Features: OpenSMILE toolkit extracting 88 low-level descriptors
- Visual Features: ResNet-50 pretrained on FER-2013, fine-tuned for facial expression recognition
- Optimizer: AdamW with learning rate 2e-5, weight decay 0.01
- Training Configuration: Batch size 16, 10 epochs, gradient clipping at norm 1.0
- Hardware: NVIDIA A100 GPU with 40GB memory
- Loss Weights: $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$ (determined via grid search)

All experiments were repeated three times with different random seeds. Early stopping was

applied based on validation F1-score with patience of 3 epochs

Discussion:

1. Impact of Instruction Conditioning:

Instruction conditioning enhances cross-modal reasoning coherence by encouraging structured evidence integration. The significant performance drop when removing instruction conditioning (7.8 percentage points) demonstrates its critical role. Analysis of generated explanations reveals that instruction-conditioned models produce more specific and evidence-grounded reasoning compared to models without this component. The approach also improves zero-shot adaptability to new emotion categories. When tested on held-out emotions not seen during training, instruction-conditioned models achieve 62.3% accuracy compared to 47.8% for standard classification models, demonstrating better generalization through structured reasoning.

2. Temporal Modeling Benefits:

Temporal modeling enables dynamic emotion tracking across conversation turns. Utterances later in conversations show greater dependency on context—the model learns to reference earlier emotional states when generating explanations. This capability proves particularly valuable for detecting emotion shifts and understanding emotional arcs in extended dialogues.

4. Qualitative Reasoning Example:

To illustrate reasoning transparency, consider the following example from the IEMOCAP dataset:

Input Utterance:"I can't believe this happened... I worked so hard."

Multimodal Observations:

- Text sentiment: Negative valence, disappointment-related keywords
- Vocal tone: Low pitch (mean F0: 142 Hz), reduced energy, slower speaking rate
- Facial expression: Downward gaze, inner brow raise, tightened lips

Model Output:

Emotion: *Sadness* (Confidence: 89.3%)

Intensity: 7.2/10

Explanation:

"The speaker exhibits negative linguistic sentiment combined with reduced vocal intensity and downward facial orientation, which collectively indicate disappointment and emotional distress. The phrase 'I can't believe' suggests unexpected negative outcomes, while the low-energy acoustic profile and facial action units AU1+AU15 are characteristic of sadness. The temporal context shows no prior anger or frustration, supporting a primary sadness classification."

This example demonstrates the model's ability to integrate cross-modal signals and generate structured, evidence-based reasoning rather than producing only a standalone label

Ethical Considerations:

Multimodal emotion inference introduces important privacy and fairness concerns that must be carefully addressed for responsible deployment.

1. Privacy Protection:

Sensitive signals such as facial expressions and vocal tone reveal personal information requiring responsible handling. Deployment in real-world systems should incorporate:

- **Data Minimization:** Collect only necessary modalities for the specific application
- **Differential Privacy:** Add calibrated noise to prevent individual identification from aggregated emotion patterns
- **Federated Learning:** Enable model training on decentralized data without

centralizing sensitive audiovisual recordings

- **Consent and Control:** Provide users clear information about data collection and options to opt out of specific modalities

2. Demographic Bias Mitigation:

Performance disparities across cultural or demographic groups must be actively mitigated. Preliminary analysis reveals accuracy variations across different accent groups (standard deviation: 4.2%) and age groups (3.8%). Mitigation strategies include:

- Balanced training data representing diverse populations
- Demographic-aware evaluation protocols
- Adversarial debiasing techniques
- Regular audits for fairness metrics across subgroups

3. Explanation Reliability:

Generated rationales may appear plausible yet be inaccurate or misleading. Verification mechanisms are necessary to ensure explanation faithfulness. Future work should incorporate human-in-the-loop validation where domain experts review model explanations for correctness.

4. Responsible Deployment:

Human oversight remains essential in sensitive domains such as healthcare and education. Emotion AI should augment rather than replace human judgment. Clear guidelines for appropriate use cases help prevent misuse in high-stakes scenarios like employment decisions or legal proceedings.

Conclusion:

This paper introduced an instruction-driven multimodal emotion reasoning framework integrating modality-aware alignment, temporal dialogue modeling, and structured explanation generation. Unlike conventional classification-based models, the proposed approach reframes emotion understanding as a conditional reasoning task guided by natural language instructions**.**

Extensive experimental evaluation demonstrates statistically significant improvements in both predictive performance (81.2% F1 on IEMOCAP) and reasoning coherence compared to state-of-the-art baselines. Ablation studies confirm that instruction conditioning provides the largest contribution to overall performance, validating the core thesis that reasoning-based approaches outperform pure classification. The framework establishes a foundation for transparent, ethical, and human-centered emotion-aware conversational AI systems. By jointly optimizing prediction accuracy and explanation quality, the model produces interpretable outputs suitable for deployment in sensitive applications requiring human understanding and trust.

Future Research Directions:

Several promising directions emerge for extending this work:

- Adaptive Instruction Learning: Meta-learning approaches to automatically discover optimal instruction formulations for different contexts
- Emotion-Aware Dialogue Policy Integration: Incorporating emotion reasoning into conversational agents for empathetic response generation
- Real-Time Parameter-Efficient Architectures: Developing compressed models suitable for edge deployment while maintaining reasoning capabilities
- Cross-Cultural Generalization: Extending the framework to handle emotion expression variations across different cultures and languages
- Multi-Task Reasoning: Expanding beyond emotion to jointly reason about

intent, sentiment, and conversational dynamics.

References:

1. Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.
2. Wu, Y., Mi, Q., & Gao, T. (2025). *A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. Biomimetics*, 10(7), Article 418.
3. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). *Multimodal transformer for unaligned multimodal language sequences*. Proceedings of ACL 2019, 6558-6569.
4. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). *Tensor fusion network for multimodal sentiment analysis.* Proceedings of EMNLP 2017, 1103-1114.
5. Zhang, S., Zhang, S., Huang, T., & Gao, W. (2018). *Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching*. IEEE Transactions on Multimedia, 20(6), 1576-1590.
6. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: *A database for facial expression, valence, and arousal computing in the wild*. IEEE Transactions on Affective Computing, 10(1), 18-31.
7. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: *Interactive emotional dyadic motion capture database. Language Resources and Evaluatio*n, 42(4), 335-359.
8. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: *A multimodal multi-party dataset for emotion recognition in conversations*. Proceedings of ACL 2019, 527-536.
9. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). *A review of affective computing: From unimodal analysis to multimodal fusion*. Information Fusion, 37, 98-125.
10. Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2005). *Multimodal approaches for emotion recognition: A survey.* Proceedings of SPIE, 5670, 56-67.
11. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). *End-to-end multimodal emotion recognition using deep neural networks.* IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301-1309.
12. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., & Marsic, I. (2018). *Multimodal affective analysis using hierarchical attention strategy with word-level alignment*. Proceedings of ACL 2018, 2225-2235.
13. Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: *An attentive RNN for emotion detection in conversations*. Proceedings of AAAI 2019, 6818-6825.
14. Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019*). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation.* Proceedings of EMNLP-IJCNLP 2019, 154-164.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need. Advances in Neural Information Processing Systems*, 30, 5998-6008.
16. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2022). *Finetuned language models are zero-shot learners.* Proceedings of ICLR 2022.
17. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2023).

Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

18. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "*Why should I trust you?" Explaining the predictions of any classifier.* Proceedings of KDD 2016, 1135-1144.
19. Lipton, Z. C. (2018). *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery*. Queue, 16(3), 31-57.
20. Camburu, O. M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). *e-SNLI: Natural language inference with natural language explanations. Advances in Neural Information* Processing Systems, 31, 9539-9549.