Multimodal Emotion Recognition in Conversation via Large Language Models and Global-Local Cross-Domain Graphs

Haobo Ma
School of Computer Science, Hubei University
Wuhan, China

Chen Huang
School of Computer Science, Hubei University
Wuhan, China

Yan Zhang
School of Computer Science, Hubei University
Wuhan, China
zhangyan@hubu.edu.cn

Chao Yang
School of Computer Science, Hubei University
Wuhan, China

Jianhua Song
School of Cyber Science and Technology, Hubei University
Wuhan, China

ABSTRACT

Multimodal Emotion Recognition in Conversation (MERC) aims to identify emotions in target utterances using multimodal data and has garnered significant interest due to its applications in conversational AI. Recognition accuracy hinges on effectively integrating multimodal cues and contextual information. However, local noise and global outliers often impair performance, while traditional approaches based on simple feature concatenation struggle to capture complex cross-modal interactions. To address these challenges, we propose LLM-EmoGraph, a novel framework that combines large language models (LLMs) with a global-local cross-domain graph architecture. Specifically, LLM-EmoGraph leverages multimodal masking strategies, a large-scale cross-domain multi-graph pretraining to improve transferability across modalities and graph structures. Then, LLM-EmoGraph further introduces an adaptive dual-scale feature fusion strategy to align semantic features across text, speech, and visual inputs. In addition, a weakly supervised hierarchical emotion classification scheme enhanced by LLMs boosts robustness and accuracy. Experiments on two benchmark datasets show that LLM-EmoGraph significantly outperforms existing methods.

KEYWORDS

Multimodal Emotion Recognition in Conversation; Large Language Models; Cross-domain Graph Architecture; Adaptive Feature Fusion; Hierarchical Emotion Classification

ACM Reference Format:
Haobo Ma, Chen Huang, Yan Zhang, Chao Yang, and Jianhua Song. 2026. Multimodal Emotion Recognition in Conversation via Large Language Models and Global-Local Cross-Domain Graphs. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HIXY4457

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HIXY4457

Figure 1: An example from the MELD dataset illustrates a multimodal conversation, where MERC is tasked with predicting the emotion label (e.g., Happy, Surprised, Fear) for each utterance.

1 INTRODUCTION

Emotions play a critical role in social life, influencing one another and reflecting internal psychological activities that guide decision-making and behavior [3]. Emotion recognition [8, 16, 17] is vital across various domains, especially in human-computer interaction, where machines must analyze users' emotional tone to offer empathetic, personalized services. Emotion Recognition in Conversation (ERC), a subfield of emotion recognition, focuses on detecting emotional expressions during conversations. With the rapid development of deep learning, ERC has been widely applied in practical scenarios like sentiment analysis [13], recommendation systems [7], and medical diagnostics [14].

In conversational settings, sequence modeling is crucial for emotion recognition, and incorporating multimodal data—such as speech, text, and facial expressions—is essential for improving performance [10], [12]. While unimodal approaches are often insufficient, multimodal emotion recognition (MERC) can leverage complementary cues across modalities. For example, in the MELD dataset, a woman's "sad" emotion is weakly conveyed through text but more clearly through visual and acoustic features. However, capturing emotional dependencies, especially in multi-party conversations, remains a challenge, as emotions are influenced by both individual states and interactions among participants [11].

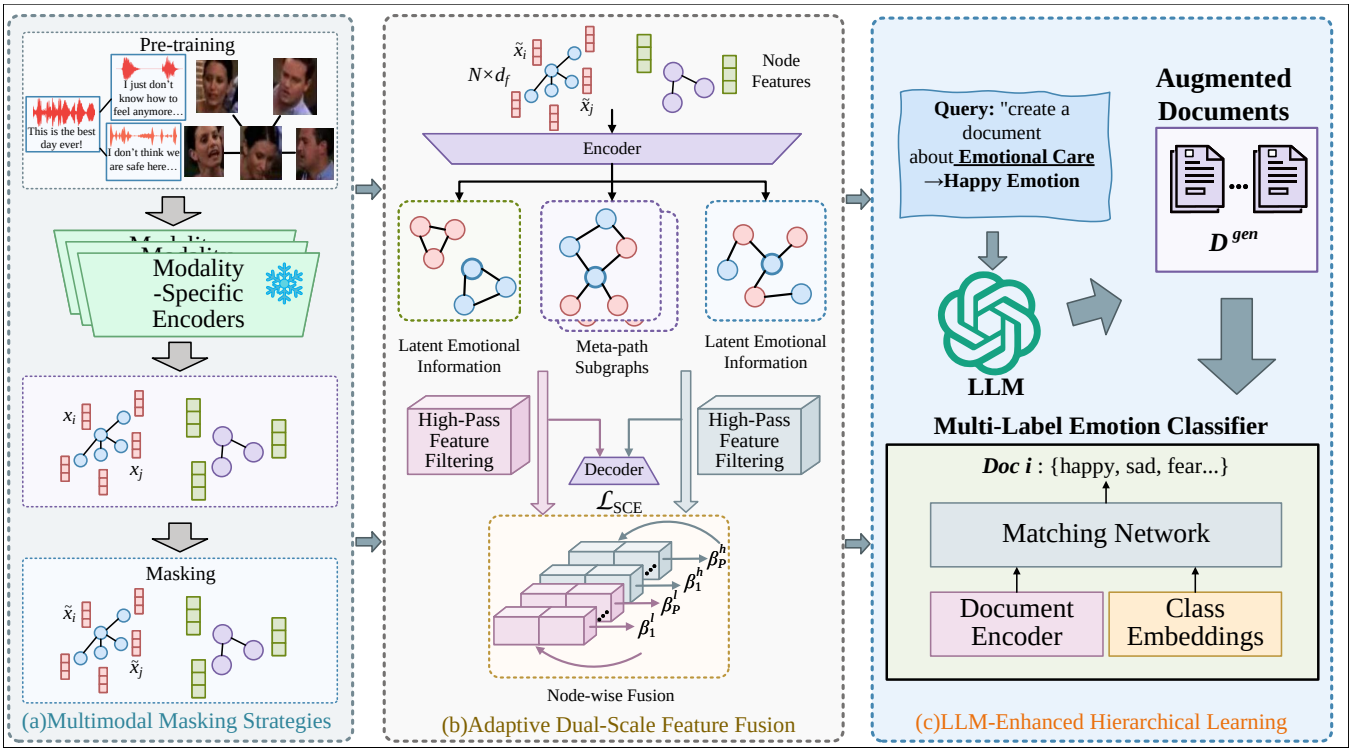

Figure 2: The overall architecture of the proposed framework LLM-EmoGraph.

Graph Neural Networks (GNNs), which excel at modeling complex relationships, have been applied in emotion recognition to improve context propagation, particularly in multimodal data. Traditional methods struggle with long-range context, while graph-based approaches, such as ConGCN [15], I-GCN [9], COGMEN [2], and MMGCN [1], improve performance by capturing contextual dependencies and integrating multimodal features. Recent studies propose further improvements, such as Li et al. [4] reducing modalities and adopting graph attention networks (GATs), GraphCFC [6] optimizing feature consistency, and GA2MIF [5] using multi-head directed graph attention networks to better capture complementary information.

Despite the progress made by graph-based methods in context propagation, several challenges remain. Firstly, existing methods generally model only specific modalities, lacking the ability for cross-modal transfer and failing to achieve functionality similar to base models without retraining or fine-tuning. Additionally, due to insufficient training of multimodal emotion recognition models, the selection of pseudo-labels is unreliable, resulting in low correlation between emotion features from different modalities and their corresponding labels, which in turn affects the accuracy of emotion recognition.

To address the aforementioned challenges, this study proposes a novel framework named LLM-EmoGraph, a multimodal emotion recognition approach based on a global-local cross-domain graph architecture powered by large language models (LLMs). Specifically, LLM-EmoGraph introduces a multimodal masking strategy and employs a large-scale cross-domain multi-graph pretraining algorithm to enable effective transfer learning across heterogeneous graph domains and modalities. Moreover, the framework incorporates an adaptive dual-scale feature fusion strategy to ensure semantic consistency across modalities, thereby effectively integrating emotional cues from text, speech, and visual inputs. Finally, LLM-EmoGraph adopts a weakly supervised hierarchical emotion recognition method enhanced by LLMs to improve the reliability of pseudo-labels, further enhancing the accuracy and robustness of emotion recognition. Through these carefully designed components, LLM-EmoGraph achieves efficient multimodal integration and significantly improves overall performance in emotion recognition tasks.

The contributions can be summarized as follows:

- We propose a cross-domain multi-graph pretraining strategy that enables LLM-EmoGraph to learn representations with uniformity and transferability across different graph domains and modalities.
- We introduce an adaptive dual-scale feature fusion strategy combined with dual-channel graph filtering, which captures both the global emotional structure and local emotional details in conversations, significantly enhancing emotion recognition performance across various dialogue patterns.
- We design a hierarchical emotion classification approach using large language models, which improves pseudo-label quality and addresses the issue of data scarcity in fine-grained classes.

REFERENCES

[1] Jian Hu, Ying Liu, Jing Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Changzhi Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). 5666–5675.

[2] Ankush Joshi, Ayush Bhat, Anshuman Jain, Ayush Vardhan Singh, and Ashutosh Modi. 2022. COGMEN: Contextualized GNN Based Multimodal Emotion Recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivana Velásquez Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 4148–4164.

[3] Gerben A. Van Kleef and Stéphane Côté. 2022. The Social Effects of Emotions. *Annual Review of Psychology* 73 (2022), 629–658.

[4] Jianhua Li, Xiaofeng Wang, Guozhen Lv, and Zhiyong Zeng. 2023. GraphMFT: A Graph Network Based Multimodal Fusion Technique for Emotion Recognition in Conversation. *Neurocomputing* 550 (2023), 126427.

[5] Jianhua Li, Xiaofeng Wang, Guozhen Lv, and Zhiyong Zeng. 2024. GA2MIF: Graph and Attention Based Two-Stage Multi-Source Information Fusion for Conversational Emotion Detection. *IEEE Transactions on Affective Computing* 15, 1 (Jan.–March 2024), 130–143.

[6] Jianhua Li, Xiaofeng Wang, Guozhen Lv, and Zhiyong Zeng. 2024. GraphCFC: A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition. *IEEE Transactions on Multimedia* 26 (2024), 77–89.

[7] Kun Liu, Yujie Wang, Qiang Li, Zhiwen Yu, and Zibin Zheng. 2023. Multimodal Graph Contrastive Learning for Multimedia-Based Recommendation. *IEEE Transactions on Multimedia* 25 (2023), 9343–9355.

[8] Wei Nie, Yubo Bao, Yuhui Zhao, and Aifeng Liu. 2024. Long Dialogue Emotion Detection Based on Commonsense Knowledge Graph Guidance. *IEEE Transactions on Multimedia* 26 (2024), 514–528.

[9] Wei Nie, Ruihua Chang, Minghao Ren, Yanan Su, and Aifeng Liu. 2022. I-GCN: Incremental Graph Convolution Network for Conversation Emotion Detection. *IEEE Transactions on Multimedia* 24 (2022), 4471–4481.

[10] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access* 7 (2019), 100943–100953.

[11] Peng Wang, Lyuba Ganushchak, Carla Welie, et al. 2024. The Dynamic Nature of Emotions in Language Learning Context: Theory, Method, and Analysis. *Educational Psychology Review* 36, 4 (2024), 105.

[12] Yutong Wang, Xianglin Zuo, Jinjie Ni, Xiaolan Weng, and Jianhua Tao. 2023. Unlocking the Power of Multimodal Learning for Emotion Recognition in Conversation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*. 5947–5955.

[13] Lingwei Wei, Deyu Hu, Wenxuan Zhou, and Shuhan Hu. 2023. Modeling Both Intra- and Intermodality Uncertainty for Multimodal Fake News Detection. *IEEE Transactions on Multimedia* 25 (2023), 7906–7916.

[14] Jin Wen, Hong Liu, Xinyao Li, Feng Wu, and Jian Yang. 2024. MsgFusion: Medical Semantic Guided Two-Branch Network for Multimodal Brain Image Fusion. *IEEE Transactions on Multimedia* 26 (2024), 944–957.

[15] Duo Zhang, Shuangyong Song, Erik Cambria, Amir Hussain, and Yang Liu. 2019. Modeling Both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-Speaker Conversations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 5415–5421.

[16] Tong Zhu, Lei Li, Jing Yang, Shuo Zhao, and Xin Xiao. 2023. Multimodal Emotion Classification with Multi-Level Semantic Reasoning Network. *IEEE Transactions on Multimedia* 25 (2023), 6868–6880.

[17] Tong Zhu, Lei Li, Jing Yang, Shuo Zhao, and Xin Xiao. 2023. Multimodal Sentiment Analysis with Image-Text Interaction Network. *IEEE Transactions on Multimedia* 25 (2023), 3375–3385.