



CemoBAM: Advancing Multimodal Emotion Recognition through Heterogeneous Graph Networks and Cross-Modal Attention Mechanisms

Nhut Minh Nguyen 

AiTA Lab,

*Dept. of Computing Fundamentals
FPT University*


Ho Chi Minh City, Vietnam
nhutnmse184534@fpt.edu.vn

Thu Thuy Le 

AiTA Lab,

*Dept. of Computing Fundamentals
FPT University*

Ho Chi Minh City, Vietnam
thuyltse170336@fpt.edu.vn

Thanh Trung Nguyen 

AiTA Lab,

*Dept. of Computing Fundamentals
FPT University*


Ho Chi Minh City, Vietnam
trungntse180355@fpt.edu.vn

Duc Tai Phan 

AiTA Lab,

*Dept. of Computing Fundamentals
FPT University*

Ho Chi Minh City, Vietnam
phantaiduc2005@gmail.com

Anh Khoa Tran 

*Modeling Evolutionary Algorithms Simulation
and Artificial Intelligence,
Faculty of Electrical and Electronics Engineering,
Ton Duc Thang University
Ho Chi Minh City, Vietnam
trananhkhoa@tdtu.edu.vn*

Duc Ngoc Minh Dang* 

AiTA Lab,

*Dept. of Computing Fundamentals
FPT University*

Ho Chi Minh City, Vietnam
ducndm2@fe.edu.vn

Abstract—Multimodal Speech Emotion Recognition (SER) offers significant advantages over unimodal approaches by integrating diverse information streams such as audio and text. However, effectively fusing these heterogeneous modalities remains a significant challenge. We propose CemoBAM, a novel dual-stream architecture that effectively integrates the Heterogeneous Graph Attention Network (CH-GAT) with the Cross-modal Convolutional Block Attention Mechanism (xCBAM). In CemoBAM architecture, the CH-GAT constructs a heterogeneous graph that models intra- and inter-modal relationships, employing multi-head attention to capture fine-grained dependencies across audio and text feature embeddings. The xCBAM enhances feature refinement through a cross-modal transformer with a modified 1D-CBAM, employing bidirectional cross-attention and channel-spatial attention to emphasize emotionally salient features. The CemoBAM architecture surpasses previous state-of-the-art (SOTA) methods by 0.32% on IEMOCAP and 3.25% on ESD datasets. Comprehensive ablation studies validate the impact of Top-K graph construction parameters, fusion strategies, and the complementary contributions of both modules. The results highlight CemoBAM's robustness and potential for advancing multimodal SER applications.

Index Terms—Multimodal emotion recognition, Speech emotion recognition, Cross-modal heterogeneous graph attention, Cross-modal convolutional block attention mechanism, Feature fusion.

I. INTRODUCTION

Speech Emotion Recognition (SER) is a crucial component of human-computer interaction, aiming to identify emotional states from spoken language. While early methods relied solely on acoustic features such as pitch and intonation, recent

advancements have shifted toward multimodal approaches that integrate audio with complementary inputs like text, facial expressions, or physiological signals to enhance robustness and accuracy. Recent studies have explored a range of fusion strategies leveraging attention mechanisms, transformer architectures, and joint representations. Guo *et al.* [1] employed spectrogram-based fusion with ResNet. Priyasad *et al.* [2] utilized Wav2Vec2 and BERT with dual memory fusion. Khan *et al.* [3] introduced a cross-modal transformer that achieved state-of-the-art (SOTA) performance. Nevertheless, effectively aligning heterogeneous modalities and capturing both intra- and inter-modal dependencies remains a key challenge in multimodal SER.

To address these challenges, graph-based and attention-based mechanisms offer promising solutions for multimodal SER. Graph structures like Graph Attention Networks (GAT) [4] capture complex relationships within and across modalities by modeling audio and text features as nodes and their similarities as edges. Meanwhile, Convolutional Block Attention Mechanisms (CBAM) [5] refine features by emphasizing emotionally salient cues through channel and spatial attention, improving the model's focus on relevant information.

In this paper, we introduce CemoBAM, which fuses audio and text modalities via two key modules: Cross-modal Heterogeneous Graph Attention Network (CH-GAT) and Cross-modal Convolutional Block Attention Mechanism (xCBAM). CH-GAT builds a heterogeneous graph with intra- and inter-modal relationships, improving cross-modal feature learning. xCBAM integrates cross-modal attention with 1D-CBAM to apply channel and spatial attention, highlighting emotionally significant features.

* Corresponding author: Duc Ngoc Minh Dang (ducndm2@fe.edu.vn)

II. METHODOLOGY

The paper proposes CemoBAM, a novel multimodal architecture designed for SER that effectively models and integrates audio and textual information. As illustrated in Fig. 1, the model comprises three main components: the feature encoding, the CH-GAT module, and the xCBAM module.

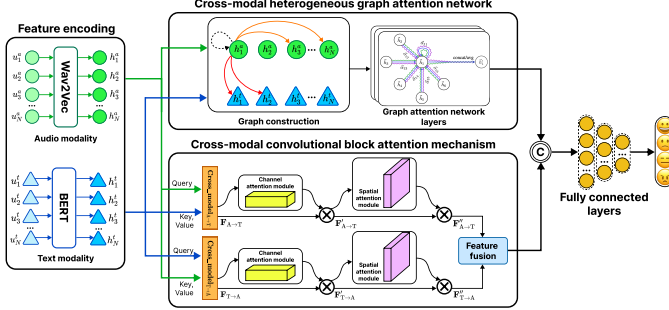


Fig. 1: The overview of the proposed CemoBAM architecture.

A. Feature encoding

We use a partially fine-tuned Wav2Vec¹ encoder for audio, with preprocessing steps including resampling, mono conversion, segmentation, and normalization to enhance emotion-specific features. For text, a BERT² encoder is used with tokenized inputs padded to a fixed length and embedded with [CLS] and [SEP] tokens. Only the upper BERT layers are fine-tuned to balance general understanding and task adaptation. This process yields modality-specific embeddings for downstream processing in the CemoBAM architecture.

B. The CH-GAT module

1) *Graph construction*: After the feature encoding, we construct a graph to capture fine-grained relational structures within and across modalities. These connections are then aggregated to form a unified edge set that defines the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For the node feature matrix (\mathcal{V}) , we concatenate the high-level embeddings derived from the text and audio modalities for each sample. This integration yields a unified representation that captures complementary information from both modalities. Formally, the resulting \mathcal{V} matrix concatenated representation ensures that each node in the graph reflects a joint multimodal context, as described in Eq. (1).

$$\mathcal{V} = [F_{\text{text}} \parallel F_{\text{audio}}], \quad (1)$$

Each node is connected to itself and its Top-K most similar neighbors in both modalities, identified via cosine similarity over L2-normalized embeddings. This reduces noise and preserves only high-confidence relationships. The edge matrix (\mathcal{E}) , which is presented in Eq. (2), captures both intra-modal similarities and implicit inter-modal interactions

through concatenated multimodal features and shared graph propagation.

$$\mathcal{E} = \bigcup_{i=1}^N \left\{ (i, i) \cup \left\{ (i, j) \mid j \in \mathcal{N}_i^{(t)} \cup \mathcal{N}_i^{(a)} \right\} \right\}, \quad (2)$$

where $\mathcal{N}_i^{(t)}$ and $\mathcal{N}_i^{(a)}$ denote the Top-K most similar neighbors of node i based on text and audio embeddings.

2) *Graph learning*: After constructing the multimodal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we apply a GAT to perform message passing and learn discriminative node-level representations. The GAT aggregates information from neighboring nodes with importance determined by learned attention coefficients. In each GAT layer, the updated node embedding \tilde{h}'_i is computed using multi-head self-attention, where representations from multiple attention heads are concatenated to enhance expressiveness. In the final GAT layer, we average the outputs of the attention heads to produce a compact and smooth node representation \tilde{h}'_i , facilitating robust modeling of intra- and inter-modal relationships for emotion recognition.

C. The xCBAM module

1) *Cross-modal transformer*: We apply a cross-attention mechanism at this stage to enable interactive feature refinement between the two modalities. The first cross-attention mechanism is audio-to-text (**Cross_model_{A→T}**), which is computed by using F_{audio} as the query and F_{text} as the key and value. This configuration allows the audio modality to selectively attend to semantically relevant text features, producing a speech-guided textual representation. In the subsequent stage, we reverse the roles in the attention mechanism to implement text-to-audio attention (**Cross_model_{T→A}**). The F_{text} serves as the query, while F_{audio} acts as the key and value. This allows the textual features to integrate corresponding acoustic cues, enhancing the semantic representation of speech.

2) *CBAM architecture*: To refine the modality-aware feature from the cross-modal transformer, we integrate a 1D-CBAM into the xCBAM module, applying channel attention followed by spatial attention.

The channel attention mechanism, described in Eq. (3), adaptively weights feature channels based on global importance. Given an input feature vector $\mathbf{F} \in \mathbb{R}^C$, average pooling and max pooling operations are applied to summarize the global context along the temporal dimension. Both pooled outputs are passed through a shared multi-layer perceptron (MLP), composed of two fully connected layers with a ReLU activation in between.

$$\mathbf{M}_c(\mathbf{F}) = \sigma \left(\mathbf{W}_2 \left(\text{ReLU}(\mathbf{W}_1 \mathbf{F}_{\text{avg}}^c) \right) + \mathbf{W}_2 \left(\text{ReLU}(\mathbf{W}_1 \mathbf{F}_{\text{max}}^c) \right) \right), \quad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{C \times C_r}$ and $\mathbf{W}_2 \in \mathbb{R}^{C_r \times C}$ are learnable weights in the shared MLP. The resulting attention weights recalibrate the input feature map via element-wise multiplication.

Spatial attention, described in Eq. (4), is applied to refine features along the sequence dimension following channel attention. The channel-wise average and max-pooled features are concatenated and passed through a 1D convolutional layer with a kernel size of 7 to capture local dependencies.

¹<https://huggingface.co/facebook/wav2vec2-base-960h>

²<https://huggingface.co/google-bert/bert-base-uncased>

$$\mathbf{M}_s(\mathbf{F}) = \sigma(f^7([\mathbf{F}_{\text{avg}}^s; \mathbf{F}_{\text{max}}^s])), \quad (4)$$

where f^7 denotes a 1D convolution operation with a kernel size of 7. This spatial attention mechanism allows the model to focus on emotionally informative regions within the sequence.

After applying the 1D-CBAM, we obtain two refined representations: $\mathbf{F}_{A \rightarrow T}''$ from the **Cross_model**_{A→T}, and $\mathbf{F}_{T \rightarrow A}''$ from the **Cross_model**_{T→A}. The 1D-CBAM module independently enhances each cross-modal stream. After cross-modal interaction, 1D-CBAM modules are applied independently to both text and audio features. This two-stage refinement enables the xCBAM module to amplify task-relevant cues in each modality before fusion, ultimately improving the model's ability to learn expressive and discriminative representations for emotion recognition.

3) *Feature fusion*: After the cross-modal transformer and the 1D-CBAM architecture, we fuse the refined features $\mathbf{F}_{A \rightarrow T}''$ and $\mathbf{F}_{T \rightarrow A}''$, which represent the audio-guided text and text-guided audio streams, respectively. To integrate these two modality-aware representations, we compare four types of fusion strategies: CLS, MIN, MAX, and MEAN. These fusion methods are evaluated to determine the most effective approach for combining multimodal information before final classification.

III. EXPERIMENT SETTINGS

A. Datasets

In this paper, we use two benchmark datasets for multimodal emotion recognition: IEMOCAP [6] and ESD [7]. The IEMOCAP dataset contains 12 hours of multimodal data with 4 emotion classes: anger, happiness, neutral, and sadness. The ESD dataset comprises 350 parallel utterances from native English and native Chinese speakers, covering 5 emotional categories: neutral, happiness, anger, sadness, and surprise. For consistency, we only use the English subset with balanced emotion classes.

B. Implementation details

All experiments were run on an NVIDIA L20 GPU. We trained CemoBAM using a learning rate of 0.0001 with Cross Entropy loss and a scheduler that decays the rate every 30 epochs. Each experiment was repeated with five random seeds, and average results are reported. Evaluation metrics include Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1-score (w-F1). Source code is available at <https://github.com/nhut-ngnn/CemoBAM>.

IV. EXPERIMENT RESULTS

A. Impact of Top-K graph construction of CH-GAT module

We examined the impact of the Top-K algorithm parameters, K_{audio} and K_{text} , which define the number of nearest neighbors retained for each node based on modality-specific cosine similarities. By evaluating 100 combinations with both values ranging from 1 to 10, we identified the optimal settings: $K_{\text{audio}} = 8$, $K_{\text{text}} = 7$ for IEMOCAP dataset and $K_{\text{audio}} = 4$

and $K_{\text{text}} = 5$ for ESD dataset. These results, illustrated in Fig. 2, proper tuning of these parameters is essential to ensure meaningful graph connectivity while minimizing noise in the heterogeneous structure.

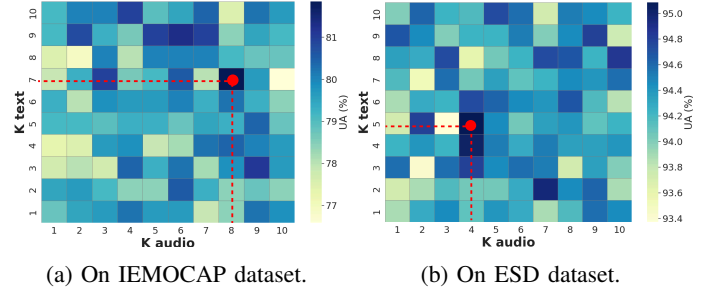


Fig. 2: Impact of Top-K graph construction of CH-GAT module on IEMOCAP and ESD datasets.

B. Impact of fusion strategies in xCBAM module

TABLE I: The impact of fusion strategies in the xCBAM module on IEMOCAP and ESD datasets.

Feature fusion	IEMOCAP			ESD		
	WA (%)	UA (%)	w-F1 (%)	WA (%)	UA (%)	w-F1 (%)
CLS	79.55	78.91	78.89	94.19	94.19	94.20
MIN	82.17	81.85	81.85	95.09	95.09	95.09
MAX	80.46	80.11	80.08	94.50	94.50	94.52
MEAN	79.50	79.17	79.19	94.48	94.48	94.50

We evaluated four fusion strategies, CLS, MIN, MAX, and MEAN, within the xCBAM module on the IEMOCAP and ESD datasets. In Table I, the MIN strategy consistently outperformed the others, achieving 82.17% WA on IEMOCAP, and 95.09% across all metrics on ESD. MIN fusion performs best because element-wise minimum fusion highlights the most salient, complementary features shared across modalities. This fusion strategies also suppresses modality-specific noise and redundant information.

C. Impact of each module in CemoBAM model

TABLE II: The impact of CH-GAT and xCBAM modules on the CemoBAM architecture.

Module	IEMOCAP			ESD		
	WA (%)	UA (%)	w-F1 (%)	WA (%)	UA (%)	w-F1 (%)
CemoBAM	82.17	81.85	81.85	95.09	95.09	95.09
- w/o xCBAM	62.49	62.32	60.07	60.27	60.27	60.17
- w/o CH-GAT	79.88	79.67	79.65	94.86	94.86	94.87

We performed an ablation study to assess the individual contributions of the CH-GAT and xCBAM modules, as presented in Table II. The full CemoBAM model achieves the highest accuracy on both datasets. When we use only the CH-GAT module, performance drops significantly, especially on ESD, where all metrics fall to approximately 60%. In contrast, using only the xCBAM module while removing CH-GAT leads to a moderate decline. These results highlight that although

xCBAM provides strong attention-based refinement, CH-GAT is critical for modeling inter- and intra-modal relationships, and both modules work synergistically to achieve optimal results.

D. Case study and comparison with SOTA

We evaluate the performance of the proposed CemoBAM model using the optimal hyperparameters identified in earlier experiments: $K_{\text{audio}} = 8$, $K_{\text{text}} = 7$ for the IEMOCAP dataset, and $K_{\text{audio}} = 4$, $K_{\text{text}} = 5$ for the ESD dataset. The CH-GAT module employs 3 GAT layers, and the xCBAM module adopts the MIN strategy for feature fusion. CemoBAM achieves SOTA performance on both datasets, with 82.17% WA, 81.85% UA, and 81.85% w-F1 on IEMOCAP, and a consistent 95.09% across all three metrics on ESD. The superior performance on ESD can be attributed to the balanced distribution of test samples across emotion classes. The comprehensive comparisons of CemoBAM with recent SOTA methods are provided in Tables III and IV. CemoBAM surpasses the previous best models by 0.32% in WA on the IEMOCAP dataset and by 3.25% in WA on the ESD dataset. These results demonstrate the model's strength in capturing cross-modal emotional representations through the combined power of graph-based reasoning and attention-based feature refinement.

TABLE III: Performance comparison of CemoBAM against SOTA multimodal emotion recognition methods on the IEMOCAP dataset.

References	Year	Modality	IEMOCAP	
			WA (%)	UA (%)
Padi <i>et al.</i> [8]	2022	Audio + Text	75.76	76.07
Priyasad <i>et al.</i> [2]	2023	Audio + Text	76.80	77.30
Pham <i>et al.</i> [9]	2023	Audio + Text	63.10	63.00
Naderi <i>et al.</i> [10]	2023	Audio	74.16	75.63
Khan <i>et al.</i> [11]	2023	Audio	72.75	-
Kyung <i>et al.</i> [12]	2024	Audio + Text	77.16	76.11
Nguyen <i>et al.</i> [13]	2024	Audio + Text	-	77.22
Khan <i>et al.</i> [3]	2025	Audio + Text	81.85	81.33
CemoBAM	2025	Audio + Text	82.17	81.85

TABLE IV: Performance comparison of CemoBAM against SOTA methods on the ESD dataset across five emotion classes. The (*) symbol denotes results evaluated on the ESD dataset with four classes, excluding "Surprise".

References	Year	Modality	ESD	
			WA (%)	UA (%)
Zhou <i>et al.</i> [7]	2022	Audio	-	89.00
Pham <i>et al.</i> [14]	2023	Audio + Text	90.47*	90.46*
Yang <i>et al.</i> [15]	2024	Audio	88.50	88.50
Khan <i>et al.</i> [3]	2025	Audio + Text	91.84	91.93
CemoBAM	2025	Audio + Text	95.09	95.09

V. CONCLUSION

This paper introduces the CemoBAM architecture, a novel multimodal SER framework that integrates audio and text through complementary CH-GAT and xCBAM modules. By

combining heterogeneous graph-based modeling and attention mechanisms to capture intra- and inter-modal relationships and emphasize emotionally salient features, CemoBAM achieves SOTA results on IEMOCAP and ESD datasets. Ablation studies confirm the synergistic effect of both modules, with MIN fusion proving most effective for feature integration. Despite these advances, challenges remain in handling noisy data and incorporating additional modalities. Future work will focus on improving robustness, developing lightweight architectures, and modeling temporal dynamics to enhance emotion recognition for human-computer interaction.

REFERENCES

- [1] Y. Guo, Y. Zhou, X. Xiong, X. Jiang, H. Tian, and Q. Zhang, "A multi-feature fusion speech emotion recognition method based on frequency band division and improved residual network," *IEEE Access*, vol. 11, pp. 86 013–86 024, 2023.
- [2] D. Prasad, T. Fernando, S. Sridharan, S. Denman, and C. Fookes, "Dual memory fusion for multimodal speech emotion recognition," in *Interspeech 2023*, 2023, pp. 4543–4547.
- [3] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "Mem-oCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion," *Scientific Reports*, vol. 15, no. 1, p. 5473, 2025.
- [4] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.
- [5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*. Berlin, Heidelberg: Springer-Verlag, 2018, p. 3–19.
- [6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [7] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [8] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 407–414.
- [9] N. T. Pham, D. N. M. Dang, B. N. H. Pham, and S. D. Nguyen, "SERVER: Multi-modal speech emotion recognition using transformer-based and vision-based embeddings," in *Proceedings of the 2023 8th International Conference on Intelligent Information Technology*, ser. ICIIT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 234–238.
- [10] N. Naderi and B. Naseri, "Cross corpus speech emotion recognition using transfer learning and attention-based fusion of Wav2Vec2 and prosody features," *Knowledge-Based Systems*, vol. 277, p. 110814, 2023.
- [11] K. Mustaqeem, A. El Saddik, F. Alotaibi, and N. Pham, "AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network," *Knowledge-Based Systems*, vol. 270, Jun. 2023.
- [12] J. Kyung, S. Heo, and J.-H. Chang, "Enhancing multimodal emotion recognition through asr error compensation and LLM fine-tuning," in *Interspeech 2024*, 2024, pp. 4683–4687.
- [13] L. H. Nguyen, N. T. Pham, M. Khan, A. Othmani, and A. El Saddik, "HuBERT-CLAP: Contrastive learning-based multimodal emotion recognition using self-alignment approach," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, ser. MMAsia '24. New York, NY, USA: Association for Computing Machinery, 2024.
- [14] N. T. Pham, L. T. Phan, D. N. M. Dang, and B. Manavalan, "SER-Fuse: An emotion recognition application utilizing multi-modal, multi-lingual, and multi-feature fusion," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, ser. SOICT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 870–877.
- [15] J. Yang, J. Liu, K. Huang, J. Xia, Z. Zhu, and H. Zhang, "Single- and cross-lingual speech emotion recognition based on WavLM domain emotion embedding," *Electronics*, vol. 13, no. 7, p. 1380, 2024.