Research Article

HemoGAT: Heterogeneous multimodal speech emotion recognition with cross-modal transformer and graph attention network

Nhut Minh NGUYEN[1], *Thanh Trung NGUYEN*[1], *Tien-Dat NGUYEN*[2], *Duc Ngoc Minh DANG*[1,*]

[1]AiTA Lab, Department of Computing Fundamental, FPT University,
D1 Street, Saigon Hi-tech Park, Tang Nhon Phu Ward, Ho Chi Minh City, 71216, Vietnam
[2] Ho Chi Minh City Radio - Television Station, Ho Chi Minh City, Vietnam

nhutnmse184534@fpt.edu.vn, trungntse180355@fpt.edu.vn, nguyentiendat0079@gmail.com, ducdnm2@fe.edu.vn

*Corresponding author: Duc Ngoc Minh Dang; ducdnm2@fe.edu.vn

DOI: 10.15598/aeee.v24i2.250415

Article history: Received Apr 27, 2025; Revised Jun 27, 2025; Accepted Jul 27, 2025; Published Jun 30, 2026.

Abstract. *Multimodal speech emotion recognition (SER) is a promising field, yet effectively fusing diverse information streams remains challenging. Addressing this requires architectures capable of modeling structural relationships across modalities with fine-grained, feature-level interactions. This paper proposes HemoGAT, a novel heterogeneous multimodal SER architecture that integrates a dual-stream architecture with two core modules: a heterogeneous multimodal graph attention network (HM-GAT) and a cross-modal transformer (CMT) to address this. The HM-GAT module captures complex structural and contextual dependencies using a heterogeneous graph constructed from deep embeddings. The CMT module enables precise cross-modal feature fusion through bidirectional cross-attention. This design effectively captures both high-level relationships and immediate cross-modal influences. HemoGAT achieves state-of-the-art (SOTA) performance on the IEMOCAP dataset and highly competitive results on the MELD dataset, demonstrating its superiority over existing methods. Extensive ablation studies were conducted to evaluate HemoGAT. We assessed the impact of the Top-K algorithm for heterogeneous graph construction and compared unimodal and multimodal fusion strategies. We also examined the contributions of the HM-GAT and CMT modules, analyzed the role of the graph attention network (GAT) in graph learning, and evaluated the effect of GAT layer depth on performance.*

Keywords

Heterogeneous graph construction, Graph attention network, Cross-modal transformer, Feature fusion, Multimodal speech emotion recognition

1. Introduction

Speech is a cornerstone modality in Human Computer Interaction (HCI), enabling more natural, intuitive, and expressive communication pathways between users and technology [1]. Its growing importance is evident in a diverse range of applications, including voice-based systems for speaker recognition [2, 3], virtual assistants [4,5], and automated dialogue systems [6]. Among these critical areas, speech emotion recognition (SER) has emerged as a particularly significant field, especially within the current era of artificial intelligence and natural language processing [7]. SER specifically addresses automatically identifying and interpreting the emotional states conveyed through acoustic signals, often analyzed with linguistic content [8]. By integrating audio signal processing and language analysis techniques, SER aims to enhance user experience in automated systems, provide sophisticated, affect-aware feedback, and ultimately improve the overall quality and empathy of human-machine interactions [9]. This capability enables wide-ranging applications such as

virtual assistants [10, 11], psychological monitoring [12], healthcare [13, 14], and education [15].

The field of SER has explored various methodologies. Traditional research often centered on unimodal systems, utilizing either audio or text as the sole input modality. For instance, Lee *et al.* [16] developed a hierarchical structure mapping speech utterances to emotion classes via successive binary classifications. Investigating feature learning within audio, Li *et al.* [17] explored hybrid deep neural networks and Hidden Markov Models (DNN-HMMs), incorporating techniques like Restricted Boltzmann Machine (RBM) -based unsupervised pre-training. On the text modality side, Batbaatar *et al.* [18] introduced the Semantic-Emotion Neural Network (SENN), designed to leverage pre-trained word representations for capturing both semantic and emotional cues. While unimodal approaches benefit from relative simplicity and lower computational cost, their significant disadvantage is the inability to leverage complementary cross-modal information. This limits performance, reduces robustness to noise or ambiguity, and potentially fails to capture the full emotional expression [19].

Consequently, multimodal SER has gained considerable traction, aiming to overcome these limitations by employing feature fusion techniques, and is now highly regarded for tackling the SER problem. Recent advancements focus on sophisticated fusion strategies while leveraging different modality combinations. Khan *et al.* [20] proposed an MSER model using audio and text, centered on a deep feature fusion technique with a multi-headed cross-attention mechanism. Ghosh *et al.* [21] proposed MMER, a multimodal multi-task framework utilizing audio and text, incorporating early fusion and cross-modal self-attention alongside auxiliary tasks to improve speech emotion recognition. He *et al.* [22] addressed computational efficiency and fusion using audio, text, and video through the domain-separated bottleneck attention (DBA) framework. To address incomplete data, Liu *et al.* [23] introduced a framework leveraging audio, text, and video, learning modality-invariant features and employing a robust imagination module with contrastive learning to handle missing modalities. The core advantage of multimodal SER lies in its potential for comprehensive understanding. However, unlocking this requires sophisticated architectures that can navigate these fusion complexities by synergistically combining methods that capture structural relationships across modalities and fine-grained feature-level interactions [24].

The proposed HemoGAT method, a heterogeneous multimodal SER with a cross-modal transformer and a graph attention network (GAT) architecture, directly confronts the limitations of prior SER techniques. Firstly, unlike unimodal approaches, which suffer from incomplete information by relying on a single data stream, HemoGAT inherently leverages both audio and text. Robust pre-processing and specialized feature encoding via partially fine-tuned transformer models ensure that rich information is extracted, overcoming the unimodal inability to capture the full emotional expression. Secondly, HemoGAT addresses the critical challenge within multimodal SER: achieving truly effective fusion. While existing multimodal methods might focus solely on direct feature interactions or broader structural patterns, HemoGAT employs an innovative dual-stream fusion strategy to capture both aspects synergistically. A heterogeneous graph built from deep embeddings enables the Heterogeneous Multimodal Graph Attention Network (HM-GAT) to model and refine complex structural and contextual relationships between and within modalities, capturing higher-level dependencies. Simultaneously, a cross-modal transformer (CMT) performs direct, fine-grained feature-level fusion through bidirectional cross-attention, enabling immediate and nuanced information exchange between audio and text.

In this paper, the main contributions are as follows:

- We propose the HemoGAT architecture for multimodal SER, leveraging HM-GAT and CMT modules to capture cross-modal interactions and effectively fuse multimodal features.
- We introduce a novel HM-GAT module to construct a new graph that effectively models node relationships, facilitating cross-modal interactions through the GAT layer. This approach enhances the integration of multimodal information, capturing intricate dependencies between audio and text features.
- We present a dual-stream fusion approach where the HM-GAT captures structural and relational dependencies via graph learning. At the same time, the CMT module performs direct, fine-grained feature-level interaction between modalities in parallel, offering complementary perspectives for enhanced multimodal representation learning.
- The HemoGAT architecture achieves state-of-the-art (SOTA) performance on the IEMOCAP dataset and highly competitive results on the MELD dataset, validating its superiority over existing multimodal SER methods.

The remainder of this paper is structured as follows: Section 2 reviews existing literature on multimodal SER, highlighting relevant advancements in feature fusion strategies, graph neural networks, and transformer-based approaches. Section 3 details the proposed HemoGAT architecture, which includes the transformer-based feature encoding for both audio and text modalities, the HM-GAT module for capturing

structural relationships, and the CMT module for fine-grained feature fusion. Section 4 presents the experimental setup, covering the datasets, implementation details, and evaluation metrics used. Section 5 discusses the performance results of HemoGAT on the benchmark datasets, comparing it with SOTA methods. Additionally, this section presents an ablation study that analyzes the individual contributions of the HM-GAT and CMT modules, the effects of graph construction parameters, the role of GAT in graph learning, and the impact of varying GAT layer depths. Finally, Section 6 summarizes the findings, reiterates the main contributions, and suggests potential future directions for research in multimodal SER.

2. Related work

This section presents a comprehensive overview of existing approaches in multimodal SER. To distinguish between different methodologies, we categorize the literature into two main groups: non-graph-based and graph-based multimodal SER. The first category comprises traditional or deep learning methods without explicitly utilizing graph representations. The second category encompasses studies that utilize graph structures to model intricate relationships among modalities, features, or temporal sequences. This division allows us to highlight the evolution of multimodal SER techniques and the unique advantages introduced by graph-based architectures.

2.1. Non-graph-based multimodal SER approaches

Prisayad *et al.* [29] proposed a novel memory-based fusion architecture for multimodal SER that surpasses traditional naive fusion approaches. Their architecture leverages explicit neural memory modules to effectively store and retrieve long-term dependencies across audio and text modalities. By integrating these modality-specific memories through cross-attentive mechanisms and compact bilinear pooling, the model learns richer and more discriminative representations, resulting in improved recognition performance during inference.

Kyung *et al.* [30] introduced a robust multimodal SER system that addresses the challenge of automatic speech recognition (ASR)-induced textual inaccuracies by integrating the ASR error compensation strategy and preference learning-based fine-tuning of a large language model (LLM). The system utilizes a cross-modal transformer to fuse speech and ASR-generated text embeddings, while the Kullback-Leibler divergence loss is employed to align ASR text with ground truth during training. Additionally, RankNet-based preference learning enhances the LLM's sensitivity to emotional subtleties. Experiments on the IEMOCAP dataset demonstrate that this approach significantly improves both weighted and unweighted accuracy, especially under high ASR error rates.

Khan *et al.* [31] introduced MemoCMT, which addresses the challenges of multimodal fusion by proposing a novel cross-modal transformer that integrates audio and textual features extracted from pre-trained HuBERT and BERT models, respectively. MemoCMT employs various aggregation techniques (e.g., CLS, mean, max, and min) to combine features, with the MIN aggregation yielding the best performance. Experiments on benchmark datasets, including IEMOCAP, ESD, and MELD, demonstrate that MemoCMT achieves SOTA unweighted and weighted accuracies, significantly advancing robust and efficient multimodal SER.

Non-graph-based multimodal SER approaches are straightforward to implement and require less computational overhead, making them suitable for systems with limited resources or real-time constraints. They effectively leverage direct modality interactions through cross-modal attention, memory-based fusion, and multi-task learning, achieving strong results when complete and high-quality data are available. However, these methods struggle to capture complex structural relationships between modalities and across utterances within conversations, which can limit their robustness under noisy or incomplete data conditions.

2.2. Graph-based multimodal SER approaches

While non-graph-based multimodal SER methods effectively capture direct modality interactions, they struggle to model complex structural relationships and contextual dependencies within and across modalities. This limits their robustness and generalization, especially under noisy or incomplete data. To address these challenges, graph-based methods offer a structured framework for modeling intra- and inter-modal dependencies, enabling robust information propagation across modalities.

Nguyen *et al.* [25] proposed CORECT, an architecture that combines a Relational Temporal Graph Convolutional Network (RT-GCN) with a Pairwise Cross-modal Feature Interaction (P-CM) module. RT-GCN effectively captures local contextual information by modeling temporal relationships between utterances across different modalities. At the same time, P-CM enhances global conversation-level understanding by integrating modality-specific features through auxiliary cross-modality interactions. Extensive experiments on benchmark datasets, including IEMOCAP and CMU-

MOSEI, demonstrated that CORECT outperforms previous SOTA methods in multimodal SER.

Chen *et al.* [26] proposed the M^{3}Net architecture, a GNN-based model that explores multivariate relationships and captures the varying importance of emotion discrepancy and commonality by leveraging multi-frequency signals. The authors enhance the ability of GNNs to model complex utterance dependencies through more effective multimodal and contextual representations. Unlike prior models that rely on pairwise interactions and tend to suppress high-frequency emotional cues, M^{3}Net introduces a dual propagation strategy: multivariate propagation via hypergraph convolution to model higher-order modality-context interactions, and multi-frequency propagation to preserve both shared and distinct emotional features across modalities.

Fan *et al.* [27] introduced a novel architecture for fusing pairwise modalities, employing a dual-channel bidirectional long short-term memory network to capture temporal features from each modality pair. These features are subsequently structured into graphs using graph convolutional networks augmented with speaker embeddings. A key innovation of their approach is a density loss that minimizes redundant information in the fused features, resulting in more distinct and comprehensive representations. Evaluations on the IEMOCAP and MELD datasets revealed that this architecture significantly outperforms existing SOTA methods, advancing multimodal SER.

In graph learning, GAT has demonstrated clear advantages in selectively capturing informative relationships, as exemplified by Nguyen *et al.* [28]. They proposed Mi-CGA, an architecture specifically designed to address incomplete multimodal emotion recognition in conversation. The approach features an Incomplete Multimodal Representation (IMR) module to simulate missing modalities, coupled with a Cross-modal GAT (CGA-Net) comprising three key components: a module for reconstructing missing features, a multi-head graph attention mechanism to enhance utterance-level representations, and a cross-modal attention module to capture inter-modal interactions effectively. Experimental evaluations on benchmark datasets demonstrate that Mi-CGA significantly surpasses several SOTA baselines, highlighting its effectiveness in the field.

Graph-based multimodal SER approaches excel at modeling both intra-modal and inter-modal relationships, utilizing graph convolution and graph attention mechanisms to capture contextual and structural dependencies within multimodal data. Despite these strengths, graph-based methods introduce higher computational and memory costs, requiring careful optimization of graph construction and GAT depth to prevent over-smoothing and overfitting. As a result, they are well-suited for applications where accuracy and structural understanding are prioritized, but practical deployment must consider resource constraints and latency.

3. Methodology

The overall architecture of HemoGAT is illustrated in Fig. 1. First, raw audio and text inputs undergo standard pre-processing. The cleaned inputs are encoded using pre-trained transformer models: Wav2Vec for audio and BERT for text, producing modality-specific embeddings. HemoGAT consists of two main components: the HM-GAT module and the CMT module. We construct a heterogeneous graph in the HM-GAT component by treating audio and text embeddings as nodes. We define a specialized edge matrix that incorporates self-node, intra-modal, and inter-modal connections to effectively capture both intra-modal and cross-modal relationships. The GAT layer then processes this graph structure, which learns to refine node representations by attending to relevant neighbors across the heterogeneous topology. Parallel to HM-GAT, the CMT module employs a two-stream multi-head cross-modality attention mechanism (**CM**), facilitating deep interaction between the audio and text modalities. This allows each modality to incorporate complementary contextual information from the other. Finally, the outputs from both HM-GAT and CMT are concatenated and passed through fully connected layers to produce the final emotion prediction. HemoGAT effectively captures structural and contextual dependencies for robust emotion recognition by integrating graph-based reasoning with cross-modality attention.

3.1. Feature encodings

Given the emotion recognition dataset (D_{emo}) containing audio and text modalities, we apply modality-specific pre-processing for feature encoding. For audio, signals are resampled to 16 kHz, converted to mono, segmented into 1-second chunks (16,000 samples), and normalized to reduce amplitude variance. For text, utterances are tokenized, padded, or truncated to a fixed length and embedded with special tokens [CLS] and [SEP] to preserve the contextual structure. The processed text is then passed through a transformer encoder to obtain deep contextual embeddings.

After pre-processing, we employ fine-tuned models for feature encoding. For the audio modality, we use Wav2Vec [32] with partial fine-tuning [33], which updates only selected layers to adapt efficiently to emotion recognition tasks. This approach leverages pre-trained acoustic representations while minimizing overfitting

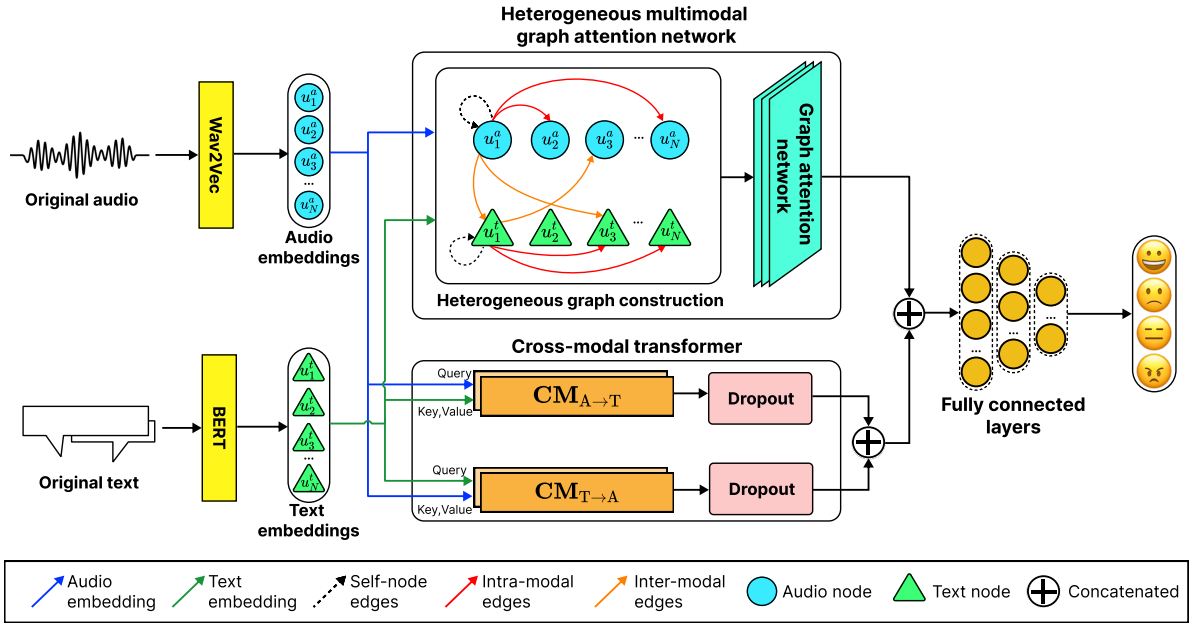


Fig. 1: Overview of the HemoGAT architecture.

and training time, yielding more robust emotion-aware audio features. On the other hand, we use BERT [34] to encode the text modality, leveraging its transformer-based architecture to capture deep semantic representations. We apply partial fine-tuning, updating only the last few layers on the emotion dataset. This strategy adapts BERT to emotion-specific patterns while reducing overfitting and training costs. The encoded feature sets for the audio and text modalities are defined in Eqs. (1) and (2).

$$F_{\text{audio}} = \{u_1^a, u_2^a, u_3^a, \ldots, u_i^a, \ldots, u_N^a\}, \tag{1}$$

$$F_{\text{text}} = \{u_1^t, u_2^t, u_3^t, \ldots, u_i^t, \ldots, u_N^t\}, \tag{2}$$

where u_i^a and u_i^t denote the feature vectors of the i-th audio and text sample, respectively. F_{audio} and F_{text} are the sets of extracted features from the audio and text modalities, obtained through Wav2Vec and BERT encoders. Here, N represents the total number of samples in the D_{emo} dataset.

3.2. Heterogeneous multimodal graph attention network (HM-GAT)

After feature encoding, the representations are passed into the HM-GAT module, a core component of the HemoGAT architecture. HM-GAT consists of two key components: heterogeneous graph construction and a stack of GAT layers, which jointly model cross-modal relationships and local structure in the multimodal data.

1) Heterogeneous graph construction

To effectively capture modality-specific relationships, we construct a heterogeneous graph that encodes local similarities in both text and audio embedding spaces. Given a dataset with N samples, each sample consists of an audio embedding (F_{audio}) and a text embedding (F_{text}). These embeddings are first L2-normalized to enable efficient computation of cosine similarity via a dot product. The resulting heterogeneous graph is denoted as $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ represents the concatenated node feature matrix containing both text and audio embeddings, and $\mathbf{E}$ represents the edge set constructed based on heterogeneous graph connections across all modalities.

In the heterogeneous graph construction process, we calculate the cosine similarity between text and audio embeddings to identify the most relevant neighbors for each node. Specifically, we select the Top-K algorithm [35] within each modality, ensuring each node retains its most informative intra-modal connections while filtering out weak associations. We select Top-K neighbors for each modality and combine them to form a multi-relational edge set that captures both intra-modal (within the same modality) and inter-modal (across different modalities) relationships. This strategy balances connectivity and sparsity, preserving meaningful relationships while ensuring computational efficiency.

Let $S^{(a)}$ and $S^{(t)}$ denote the cosine similarity matrices for the audio and text modalities, respectively. The cosine similarity between nodes is computed as Eqs. (3), and (4), where i, j are the different nodes in the graph.

$$S_{i,j\neq i}^{(a)} = F_{\text{audio}} \cdot (F_{\text{audio}})^{\top}, \tag{3}$$

$$S_{i,j\neq i}^{(t)} = F_{\text{text}} \cdot (F_{\text{text}})^{\top}. \tag{4}$$

For each node i, we select the Top-K algorithm based on cosine similarity in the text and audio modalities. This selection process ensures that each node establishes strong intra-modal and inter-modal connections. This allows the model to capture meaningful local structures within each modality while preserving cross-modal relationships. The Top-K selection criteria for each modality are formally defined as Eqs. (5) and (6).

$$\mathcal{N}_i^{(a)} = \text{Top-}k_{\text{audio}}\left(S_{i,j\neq i}^{(a)}\right), \tag{5}$$

$$\mathcal{N}_i^{(t)} = \text{Top-}k_{\text{text}}\left(S_{i,j\neq i}^{(t)}\right). \tag{6}$$

Once the Top-K neighbors are identified for both modalities, we integrate them into a unified heterogeneous graph structure. This enables information to propagate across both intra-modal and inter-modal connections, enriching node representations through graph-based message passing. The resulting graph effectively models the multimodal relationships between text and audio, forming a robust foundation for downstream tasks such as emotion recognition.

The final heterogeneous edge set $\mathbf{E}$ is constructed by combining these nearest neighbors and including self-loops, ensuring each node maintains both intra-modal and inter-modal connections. The complete edge set is defined as Eq. (7).

$$\mathbf{E} = \bigcup_{i=1}^{N} \left\{ (i,i) \cup \left\{ (i,j) \mid j \in \mathcal{N}_i^{(t)} \cup \mathcal{N}_i^{(a)} \right\} \right\}. \tag{7}$$

The node feature matrix $\mathbf{V}$ is formed by concatenating the feature embeddings from both text and audio modalities. This concatenation ensures that each node incorporates information from both modalities, resulting in a feature matrix of dimension $\mathbf{V} \in \mathbb{R}^{N\times(d_{\text{audio}}+d_{\text{text}})}$, where N is the number of samples, and d_{audio} and d_{text} represent the embedding dimensions of the audio and text modalities, respectively. The node representation is expressed as Eq. (8).

$$\mathbf{V} = [F_{\text{text}} \,\|\, F_{\text{audio}}]. \tag{8}$$

2) Graph learning

Attention mechanisms have recently gained prominence due to their significant contributions to modeling interactions between entities. The GAT [36] introduces an attention-based approach for graph learning, allowing each node to selectively aggregate information from its most relevant neighbors. Unlike traditional Graph Convolutional Networks (GCNs), which apply uniform weighting to all neighbors, GAT dynamically assigns learnable attention scores, ensuring that more influential relationships contribute more significantly to the learned representations [37]. Attention-based graph learning is essential in multimodal learning, where speech and text modalities exhibit complex interdependencies. The HM-GAT extends GAT principles to a multimodal setting, enabling both intra-modal (within text or audio) and inter-modal (between text and audio) relationships to be effectively captured. This allows the model to preserve local semantic structures while facilitating cross-modal information exchange.

We employ a three-layer GAT architecture that progressively refines node representations to model multimodal relationships. The first two GAT layers utilize concatenation-based multi-head attention, preserving diverse feature information from neighboring nodes and allowing the model to capture fine-grained intra-modal and inter-modal dependencies. This step enhances the richness of learned embeddings by aggregating contextual information across multiple attention heads, which is visualized in Eq. (9).

$$\tilde{h}_i' = \Big\|_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k W^k \tilde{h}_j\right), \tag{9}$$

where α_{ij}^k denotes the attention coefficient computed by the k-th attention head, W^k is the transformation weight matrix, and $\sigma(\cdot)$ is the softmax function applied across attention weights.

In the final GAT layer, we transition to mean-based aggregation in Eq. (10), which compacts the multi-head representations into a unified embedding.

$$\tilde{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k W^k \tilde{h}_j\right). \tag{10}$$

This shift ensures a more stable and generalizable node representation, reducing redundancy while preserving essential information. The hierarchical design of our HM-GAT module enables a balance between local feature retention in the early layers and global generalization in the final layer. We enhance multimodal feature fusion by integrating this structured graph learning approach, improving SER performance.

3.3. Cross-modal transformer (CMT)

Parallel to the HM-GAT module, which utilizes feature embeddings to construct a heterogeneous graph

for relational modeling, the CMT module focuses on direct feature interaction between modalities at the embedding level. This module facilitates the exchange of information between speech and text modalities through a bidirectional cross-attention mechanism. These feature embeddings serve as input for the CMT, where cross-attention is applied to refine each representation in modality based on information from the other. This process ensures that speech features gain contextual understanding from text, and vice versa, before being integrated into the graph structure.

In the first stage, audio-to-text attention is computed by treating F_{audio} as the query and F_{text} as the key-value pair. This operation allows the speech modality to attend to relevant text features, producing an enhanced representation of speech-influenced text. The function of audio-to-text attention is visualized in Eq. (11).

$$\mathbf{CM}_{\text{A}\to\text{T}} = \sigma\left(\frac{F_{\text{audio}}\mathbf{W}_{Q_a}\left(\mathbf{W}_{K_t}\right)^{\top}(F_{\text{text}})^{\top}}{\sqrt{d_k}}\right)F_{\text{text}}\mathbf{W}_{V_t}, \tag{11}$$

where $\mathbf{W}_{Q_a}$, $\mathbf{W}_{K_t}$, and $\mathbf{W}_{V_t}$ are trainable weight matrices that project the audio and text embeddings into query, key, and value representations, respectively. The term d_k represents the key vector dimension, ensuring proper scaling in the attention computation.

Similarly, text-to-audio attention is performed by using F_{text} as the query and F_{audio} as the key-value pair, enabling the text modality to incorporate relevant information from speech. The role of text-to-audio attention is illustrated in Eq. (12).

$$\mathbf{CM}_{\text{T}\to\text{A}} = \sigma\left(\frac{F_{\text{text}}\mathbf{W}_{Q_t}(\mathbf{W}_{K_a})^{\top}(F_{\text{audio}})^{\top}}{\sqrt{d_k}}\right)F_{\text{audio}}\mathbf{W}_{V_a}, \tag{12}$$

where $\mathbf{W}_{Q_t}$, $\mathbf{W}_{K_a}$, and $\mathbf{W}_{V_a}$ are the query, key, and value projection matrices for the text and audio modalities, respectively. The cross-attention mechanism ensures that each modality refines its representation by attending to the most relevant features from the other modality.

The final CMT representation feature (F_{CMT}) is obtained by concatenating the attended feature vectors from both directions, resulting in a unified representation of size $F_{\text{CMT}} \in \mathbb{R}^{N\times 2*(d_{\text{audio}}+d_{\text{text}})}$. The operation in Eq. (13) captures complementary information from speech and text, enhancing the model's ability to recognize emotions. By leveraging the cross-attention mechanism, CMT effectively integrates multimodal information while preserving modality-specific characteristics, leading to more robust feature learning for speech emotion recognition.

$$F_{\text{CMT}} = \left[\mathbf{CM}_{\text{A}\to\text{T}} \,\|\, \mathbf{CM}_{\text{T}\to\text{A}}\right]. \tag{13}$$

4. Experimental setup

4.1. Datasets

This study utilizes two publicly available real-life datasets for the multimodal SER task: IEMOCAP [38] and MELD [39]. Detailed statistics of these datasets are provided in Tab. 1.

The IEMOCAP dataset comprises 12 hours of audiovisual recordings from 10 speakers, collected at the SAIL Lab at the University of Southern California. It includes video, speech, facial motion capture, and text transcriptions. In this study, we categorized the data into four emotion classes: anger (1,103 utterances), happiness (1,635 utterances), neutral (1,708 utterances), and sadness (1,084 utterances) for emotion classification tasks.

The MELD dataset, an extension of EmotionLines, enriches textual data with audio and visual modalities. It contains over 1,400 dialogues and 13,000 utterances from the Friends TV series, labeled with seven emotions and sentiment classes. The emotion classes contain: anger, disgust, sadness, joy, neutrality, surprise, and fear. MELD serves as a comprehensive multimodal benchmark for emotion and sentiment analysis.

4.2. Implementation details

All experiments are implemented using the PyTorch framework and are conducted on an NVIDIA Tesla P100 GPU to ensure efficient training and evaluation. A learning rate scheduler reduces the learning rate by a factor of 0.1 every 30 epochs if the validation performance plateaus, promoting stable convergence. To ensure robustness and reproducibility, we repeat each experiment using five different random seeds, and the final results are reported as the mean across runs. The cross-entropy loss [40] is used to optimize the model for multi-class emotion classification, defined as Eq. (14).

$$L = -\sum_{i=1}^{N_c} y_i \log(p_i), \tag{14}$$

where N_c is the total number of emotion classes, y_i is the one-hot encoded ground-truth label, and p_i is the predicted probability for class i. Model checkpointing is applied to save the best-performing models for further evaluation and deployment. The source code of the HemoGAT architecture is publicly available at `https://github.com/nhut-ngnn/HemoGAT`.

To evaluate the HemoGAT architecture, we use accuracy (Acc) and the weighted F1-score (w-F1), which provide a balanced measure of classification performance

Tab. 1: Summary of experimental datasets.

Datasets	Classes	Utterances			Number of speakers	Total time
		Train	Valid	Test		
IEMOCAP	4	4,407	551	551	10 speakers	12 h
MELD	7	9,988	1,109	2,611	Multiple speakers	13.67 h

across emotion classes. The equations of the two metrics are provided in Eqs. (15), and (16).

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{15}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

$$\text{w-F1} = \frac{1}{N_c} \sum_{i=1}^{N} w_i \cdot F1_i, \tag{16}$$

where N_c is the number of classes, $F1_i$ is the F1-score for class i, and w_i is the corresponding class weight based on its support.

4.3. Baseline models

We compare HemoGAT with SOTA baselines across non-graph and graph neural network models for each dataset. On IEMOCAP, non-graph models include Satoso *et al.* [41] with self-attention weight correction, Zhao *et al.* [42] using prompt-based pre-training, Prisayad *et al.* [29] with dual memory fusion, and Kyung *et al.* [30] incorporating automatic speech recognition error compensation with large language model fine-tuning. Nguyen *et al.* [43] applied contrastive self-alignment. Ma *et al.* [44] proposed the Emotion2vec framework, which is pre-trained on open-source, unlabeled emotion data through self-supervised online distillation. Khan *et al.* [31] explored cross-modal transformer-based fusion. For graph-based models, on the IEMOCAP dataset, Nguyen *et al.* [28] proposed a cross-modal GAT, Qi *et al.* [45] proposed an innovative multimodal fusion graph convolutional network (MFGCN). While on MELD, Li *et al.* [46] introduced GraphMFT for multimodal fusion, Li *et al.* [47] proposed GraphCFC, a directed graph-based cross-modal feature complementation, and Lu *et al.* [48] utilized bi-stream graph learning.

5. Results and discussions

5.1. Performance results

We evaluate HemoGAT on the IEMOCAP and MELD datasets using a three-layer GAT architecture, employing the Top-K algorithm to optimize graph construction with dataset-specific k_{audio} and k_{text} values. These results of the HemoGAT model are shown in Tab. 2. On IEMOCAP, the best performance is achieved with $k_{\text{audio}} = 4$ and $k_{\text{text}} = 5$, yielding an accuracy of 81.97% and a weighted F1-score (w-F1) of 82.18%. On MELD, using $k_{\text{audio}} = 8$ and $k_{\text{text}} = 2$, the model attains an accuracy of 63.29% and a w-F1 of 60.73%. To further validate robustness, we report the standard deviations and 95% confidence intervals (CIs) across five independent runs, with IEMOCAP showing a standard deviation of ± 0.37 and 95% CI of ± 0.97 for accuracy, and a standard deviation of ± 0.39 and 95% CI of ± 1.00 for w-F1. On MELD, the model achieves a standard deviation of ± 0.62 and 95% CI of ± 1.00 for accuracy, and a standard deviation of ± 0.66 and 95% CI of ± 0.85 for w-F1. Additionally, we report Precision and Recall metrics to provide a more comprehensive evaluation, where HemoGAT achieves 81.85% precision and 81.71% recall on IEMOCAP, and 60.21% precision and 62.90% recall on MELD.

To provide a comprehensive evaluation under imbalanced emotion distributions, we further visualize the Area Under the Receiver Operating Characteristic (AUC-ROC) curves and the Precision-Recall (PR) curves of HemoGAT on both datasets, as shown in Fig. 2 and Fig. 3. These curves illustrate the model's ability to effectively discriminate between emotion classes while maintaining a high precision-recall balance across different thresholds. On IEMOCAP, the curves demonstrate consistently high separability across classes, while on MELD, the curves indicate competitive discrimination performance despite the dataset's conversational complexity and class imbalance. These results validate the generalization capability of HemoGAT across different datasets and emotion categories, highlighting its effectiveness in multimodal speech emotion recognition.

Tab. 3 compares HemoGAT with other SOTA multimodal SER methods, highlighting its effectiveness in capturing multimodal dependencies. On IEMOCAP,

Tab. 2: The performance results of HemoGAT architecture on IEMOCAP and MELD datasets.

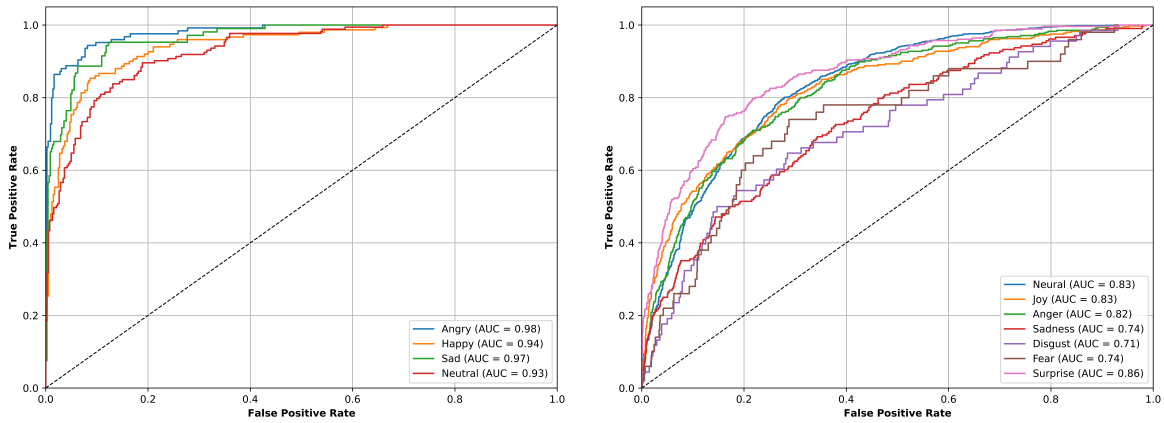
Dataset	Value	Performance metrics			
		Acc (%)	w-F1 (%)	Precision (%)	Recall (%)
IEMOCAP	**Mean**	**81.97**	**82.18**	**81.85**	**81.71**
	Standard deviation	± 0.37	± 0.39	± 0.41	± 0.37
	95% CIs	± 0.97	± 1.00	± 0.86	± 0.78
MELD	**Mean**	**63.29**	**60.73**	**60.21**	**62.90**
	Standard deviation	± 0.62	± 0.66	± 0.77	± 0.62
	95% CIs	± 1.00	± 0.85	± 1.14	± 1.00

Tab. 3: Performance comparison of multimodal SER methods. **Bold** denotes the best result, and underline denotes the second-best. In the modalities column, A = audio, T = text, V = visual.

References	Year	Modalities	GAT	IEMOCAP		MELD	
				Acc (%)	w-F1 (%)	Acc (%)	w-F1 (%)
Non-graph-based approaches							
Santoso *et al.* [41]	2022	A + T	✗	76.8	76.6	-	-
Zhao *et al.* [42]	2022	A + T + V	✗	80.01	81.09	-	-
Prisayad *et al.* [29]	2023	A + T	✗	76.8	77.3	-	-
Kyung *et al.* [30]	2024	A + T	✗	76.11	77.16	-	-
Nguyen *et al.* [43]	2024	A + T	✗	77.22	-	-	-
Ma *et al.* [44]	2024	A + T	✗	81.68	80.75	-	-
Khan *et al.* [31]	2025	A + T	✗	81.33	81.85	**64.18**	**62.52**
Graph-based approaches							
Li *et al.* [46]	2023	A + T + V	✓	-	-	61.30	58.37
Lu *et al.* [48]	2024	A + T + V	✓	-	-	62.76	59.21
Li *et al.* [47]	2024	A + T	✓	-	-	59.96	57.46
Nguyen *et al.* [28]	2025	A + T	✓	81.50	-	-	-
Qi *et al.* [45]	2025	A + T	✗	78.2	78.3	-	-
HemoGAT (Ours)		A + T	✓	**81.97**	**82.18**	63.29	60.73

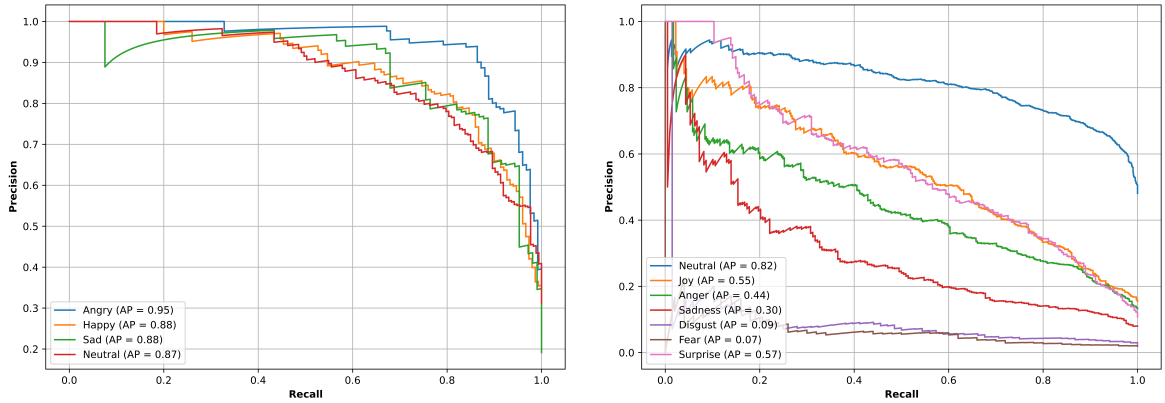
HemoGAT achieves the best performance, surpassing previous SOTA models such as the approach of Ma *et al.* [44] and Khan *et al.* [31], with improvements of 0.29% in accuracy and 0.33% in w-F1. On MELD, HemoGAT secures the second-best performance, achieving 63.29% accuracy and a w-F1 of 60.73%, closely trailing with Khan *et al.* [31], which attains 64.18% and 62.52%, respectively. Tab. 3 also includes GAT indicator columns to clarify whether each method leverages graph attention mechanisms, emphasizing the role of graph-based learning in achieving these competitive results. These results confirm the robustness and generalizability of HemoGAT across datasets, demonstrating its competitive performance on MELD while consistently achieving SOTA results on IEMOCAP. The consistent improvements in both accuracy and weighted F1 highlight HemoGAT's capability in effectively capturing and leveraging multimodal emotional cues.

HemoGAT's dual-stream architecture excels on the IEMOCAP dataset, effectively capturing its dyadic interactions and four emotion classes through HM-GAT and CMT modules. The Top-K algorithm plays a critical role by selecting optimal neighbors, ensuring the graph retains the most informative connections, reducing noise, and enhancing the HM-GAT's ability to model emotional patterns. By contrast, the optimal configuration for the MELD dataset prioritizes audio connections, emphasizing prosodic cues in noisy conversational contexts. Although HemoGAT ranks second on MELD, it can outperform other graph-based methods, such as GraphMFT [47] (61.30% Acc) and bi-stream graph [48] (62.76% Acc), underscoring the strength of its hybrid approach. Further analysis indicates that modality conflicts may persist under certain conditions, likely due to the reliance on direct concatenation for fusion. Investigating advanced fusion techniques that dynamically adjust the contributions of different modalities could significantly improve robustness. The ablation study further confirms that the three-layer GAT architecture strikes an optimal balance, capturing higher-order dependencies without overfitting, as deeper layers lead to performance degradation due to over-smoothing.



(a) On IEMOCAP dataset. (b) On MELD dataset.

Fig. 2: AUC-ROC curves of the HemoGAT architecture on both IEMOCAP and MELD datasets.



(a) On IEMOCAP dataset. (b) On MELD dataset.

Fig. 3: PR curves of the HemoGAT architecture on both IEMOCAP and MELD datasets.

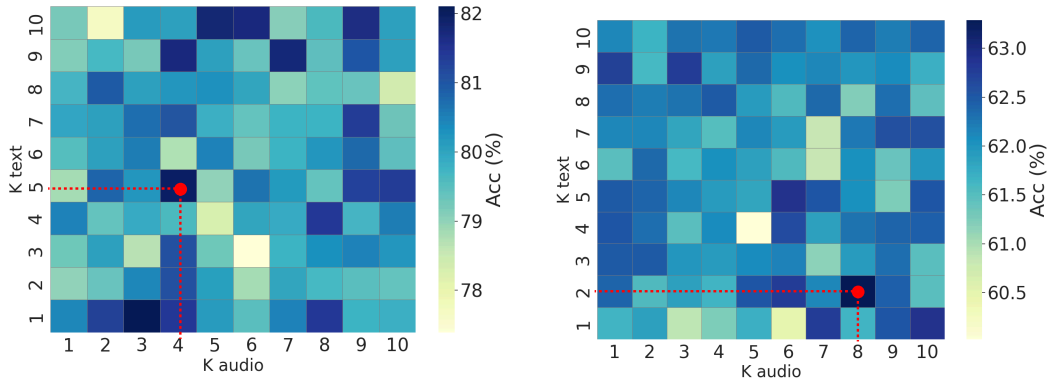
5.2. Ablation study

1) Impact of the Top-K algorithm on heterogeneous graph construction

To optimize the construction of the heterogeneous graph, we rigorously evaluated the influence of the Top-K algorithm parameters, k_{audio} and k_{text}, which govern the number of nearest neighbors retained for each node based on modality-specific cosine similarities. A comprehensive comparison was performed, testing 100 distinct configurations for each dataset by systematically varying both k_{audio} and k_{text} across the range of 1 to 10. As definitively illustrated by the accuracy heatmaps in Fig. 4, this exhaustive grid search identified the optimal hyperparameters achieving peak performance: $k_{\text{audio}} = 4$ and $k_{\text{text}} = 5$ for IEMOCAP, and $k_{\text{audio}} = 8$ and $k_{\text{text}} = 2$ for MELD. These results highlight the critical need for dataset-specific tuning of these parameters. The careful selection of K values is fundamental, as it precisely controls the graph's edge structure, striking a crucial balance between incorporating sufficient connectivity to capture relevant intra- and inter-modal relationships and maintaining sparsity to mitigate noise and irrelevant connections. This optimized graph topology is essential for practical information propagation within the subsequent GAT layers and is integral to HemoGAT's demonstrated efficacy.

2) Impact of modules and modalities on HemoGAT architecture

The effectiveness of HemoGAT is evaluated in Tab. 4 from two perspectives: modality evaluation and module evaluation. We assess the performance of each modality (audio-only and text-only) and their fusion (audio + text). The results demonstrate that combining both modalities consistently improves performance across both datasets. On IEMOCAP, audio-only and text-only models achieve 76.30% and 68.52% accuracy, respectively, whereas the fusion model attains 81.97%



(a) On IEMOCAP dataset. (b) On MELD dataset.

Fig. 4: Grid search heatmaps for optimal k_{audio} and k_{text} in graph construction. The red dot highlights the best-performing configuration on each dataset.

Tab. 4: Comparison of HM-GAT and CMT module efficiency on the IEMOCAP and MELD datasets.

Datasets	Modules	Audio		Text		Audio and text	
		Acc (%)	w-F1(%)	Acc (%)	w-F1(%)	Acc (%)	w-F1(%)
IEMOCAP	w/o HM-GAT	70.60	70.54	65.70	65.74	77.86	77.79
	w/o CMT	58.26	58.28	53.01	51.97	62.81	60.63
	HemoGAT	**76.30**	**76.79**	**68.52**	**68.76**	**81.97**	**82.18**
MELD	w/o HM-GAT	48.95	40.33	62.44	58.40	62.09	59.93
	w/o CMT	49.67	39.62	58.14	53.75	55.13	45.47
	HemoGAT	**49.90**	**40.71**	**62.55**	**59.49**	**63.29**	**60.73**

accuracy and 82.18% w-F1. A similar trend is observed on MELD, where multimodal fusion improves performance, reaching 63.29% accuracy and 60.73% w-F1. These results confirm that integrating both modalities provides richer emotional cues, enhances recognition accuracy.

To analyze the contributions of the HM-GAT and CMT modules, we conduct ablation experiments by removing each component. On IEMOCAP, eliminating HM-GAT reduces accuracy by 4.11% and w-F1 by 4.39%, while removing CMT causes a larger drop of 19.16% in accuracy and 21.55% in w-F1. On MELD, removing HM-GAT results in a 1.20% accuracy decrease and a 0.80% w-F1 drop, whereas excluding CMT significantly lowers accuracy by 8.16% and w-F1 by 15.26%. These findings highlight that HM-GAT effectively enhances cross-modal interactions while CMT improves feature fusion, both playing crucial roles in achieving SOTA performance.

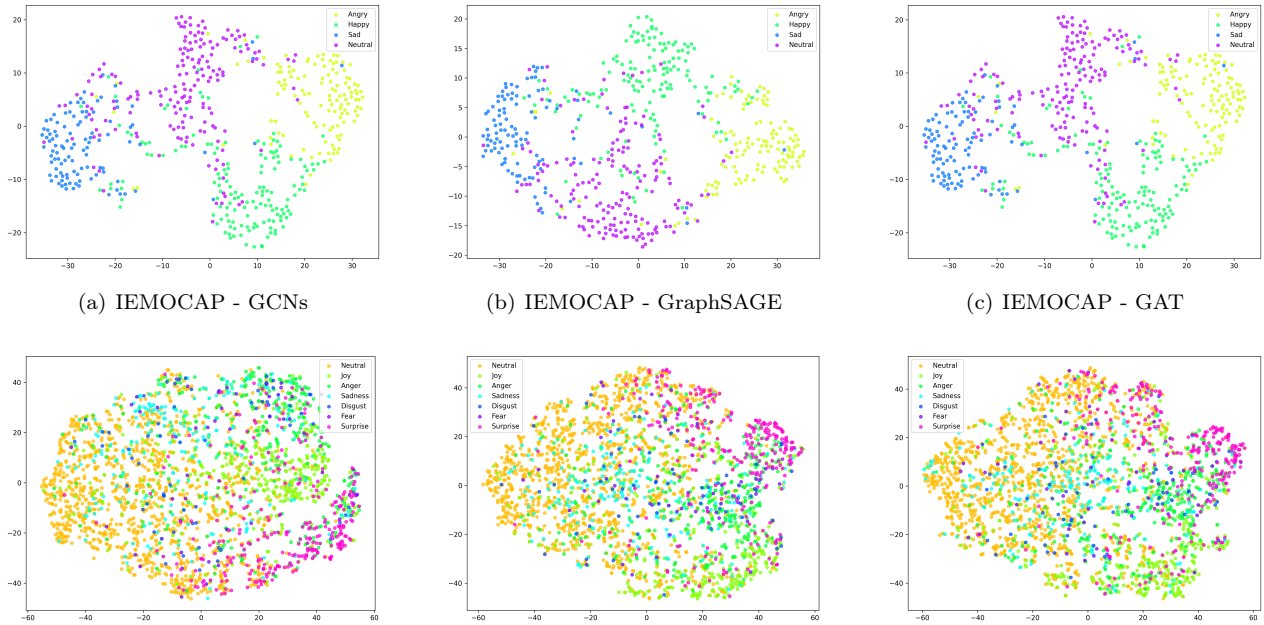
These findings highlight that HM-GAT plays a crucial role in modeling node relationships, enhancing cross-modal interactions, and effectively clustering similar emotional patterns. By leveraging graph attention mechanisms, HM-GAT strengthens meaningful connections in the multimodal feature space, improving representation learning. On the other hand, CMT is responsible for feature integration, effectively utilizing and fusing audio and text features. The CMT component refines feature representations by aligning modality-specific characteristics, ensuring that complementary information is preserved, and contributing to robust emotion recognition. HM-GAT and CMT significantly enhance multimodal fusion, enabling HemoGAT to achieve SOTA performance.

3) Impact of graph learning methods on HemoGAT architecture

To evaluate the impact of different graph learning strategies within HemoGAT, we conducted a comparative analysis of GAT, GCNs [49], and GraphSAGE [50]. We trained the model using each graph learning method under the same experimental settings and visualized the learned embeddings using t-SNE in Fig. 5, and compared the Acc, w-F1, parameter count, and floating-point operations (FLOPs) in Tab. 5. This systematic comparison enables us to evaluate how each method captures multimodal emotional patterns and the asso-

Tab. 5: Comparison of performance and computational cost of graph learning methods on IEMOCAP and MELD datasets.

Datasets	Methods	Performance results		Computational costs	
		Acc (%) ↑	w-F1 (%) ↑	Params ↓	FLOPs ↓
IEMOCAP	GCNs	81.58	81.57	**201.09M**	**20.68G**
	GraphSAGE	81.30	81.32	202.14M	21.26G
	GAT	**81.97**	**82.18**	207.40M	24.17G
MELD	GCNs	61.82	58.71	**201.09M**	**31.49G**
	GraphSAGE	61.69	58.54	202.14M	34.23G
	GAT	**63.29**	**60.73**	207.40M	47.93G



(a) IEMOCAP - GCNs (b) IEMOCAP - GraphSAGE (c) IEMOCAP - GAT

(d) MELD - GCNs (e) MELD - GraphSAGE (f) MELD - GAT

Fig. 5: t-SNE visualization of learned embeddings on the IEMOCAP and MELD datasets using different graph learning methods.

ciated computational trade-offs within the HemoGAT architecture.

While GCNs and GraphSAGE effectively capture structural relationships by aggregating neighborhood information, their fixed aggregation schemes limit the ability to prioritize informative neighbors, reducing their capacity to model nuanced emotional patterns. As a result, they achieve lower performance, with GCNs obtaining 81.58% Acc and 81.57% w-F1 on IEMOCAP and 61.82% Acc with 58.71% w-F1 on MELD. In contrast, GAT employs learnable attention to dynamically weight neighbors, enabling it to capture fine-grained, modality-specific dependencies, leading to higher performance on IEMOCAP (81.97% Acc, 82.18% w-F1) and MELD (63.29% Acc, 60.73% w-F1). However, this improvement comes at the cost of increased computational resources, with GAT requiring 207M parameters and 24.17G FLOPs on IEMOCAP, which is higher than those of GCNs and GraphSAGE.

This ablation study demonstrates that GAT offers a clear advantage in terms of improved accuracy and the ability to capture detailed emotional patterns for multimodal emotion recognition, making it highly effective for high-performance applications where accuracy is critical. However, the primary disadvantage of GAT is its higher computational and memory cost, which may limit its applicability in real-time or resource-constrained environments. Therefore, practitioners should consider the trade-off between the performance improvements offered by GAT and its computational demands when deploying multimodal affective computing systems in practical scenarios.

4) Impact of number of GAT layers on graph learning

We systematically evaluated configurations ranging from 1 to 4 layers to ascertain the optimal GAT depth,

Tab. 6: Comparison of the efficiency of GAT layers on IEMOCAP and MELD datasets.

Layers	IEMOCAP		MELD	
	Acc (%)	w-F1(%)	Acc (%)	w-F1(%)
1	79.77	79.74	61.40	58.05
2	80.05	80.23	61.69	58.27
3	**81.97**	**82.18**	**63.29**	**60.73**
4	80.16	80.20	62.04	59.85

holding other architectural elements constant. The empirical results in Tab. 6 for both the IEMOCAP and MELD datasets demonstrate a clear advantage for a 3-layer architecture, achieving 81.97% Acc and 82.18% w-F1 on IEMOCAP, and 63.29% Acc and 60.73% w-F1 on MELD. Increasing the depth to this point significantly enhances the model's ability to capture intricate, higher-order nodal dependencies and interactions within the heterogeneous graph, progressively refining representations by aggregating information from expanding neighborhoods. However, extending the depth to 4 layers leads to performance degradation due to overfitting and pronounced oversmoothing, resulting in reduced performance with 80.16% Acc and 80.20% w-F1 on IEMOCAP, and 62.04% Acc and 59.85% w-F1 on MELD. Oversmoothing occurs when excessive message passing in deeper GAT layers causes node representations to converge toward similar values, eroding feature distinctiveness and diminishing the model's ability to differentiate emotional patterns, as evidenced by these lower scores. Consequently, a 3-layer GAT configuration is unequivocally identified as the optimal choice for HemoGAT, providing the most effective balance between representational capacity and robust generalization.

6. Conclusion

In this paper, we propose HemoGAT, a novel architecture for multimodal SER designed to address the limitations of unimodal systems and the fusion challenges in existing multimodal approaches. By synergistically integrating the HM-GAT module for capturing structural relationships and the CMT module for fine-grained feature interactions, HemoGAT achieves a comprehensive and robust fusion of audio and text modalities. Our experiments demonstrate HemoGAT's effectiveness in achieving SOTA performance on the IEMOCAP dataset and highly competitive results on the MELD dataset. Ablation studies further validate the significant contributions of both core modules and underscore the importance of optimal hyperparameter configuration, confirming HemoGAT as a substantial advancement in leveraging graph networks and transformers for sophisticated affective computing. Future work could explore the incorporation of additional modalities or the adaptation of the model for enhanced real-world robustness.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Nhut Minh Nguyen, Thanh Trung Nguyen, and Duc Ngoc Minh Dang conceived and planned the research project. Duc Ngoc Minh Dang developed the overarching research goals and aims, while Nhut Minh Nguyen and Thanh Trung Nguyen designed the methodology and created the models. Nhut Minh Nguyen and Thanh Trung Nguyen developed the software, implementing the computer code and supporting algorithms. Nhut Minh Nguyen prepared the visualizations and data presentations. Nhut Minh Nguyen and Thanh Trung Nguyen wrote the initial draft of the manuscript, with Duc Ngoc Minh Dang providing critical review, proofreading, and editing. Duc Ngoc Minh Dang validated the reproducibility of the results and managed the project, overseeing its planning and execution. All authors discussed the findings and contributed to the final manuscript.

References

[1] DENG, L., A. ACERO, Y. WANG, K. WANG, H. HON, J. DROPPO, M. MAHAJAN, and X. HUANG. A speech-centric perspective for human-computer interface. *2002 IEEE Workshop on Multimedia Signal Processing*. 2002, pp. 263–267. ISBN 0-7803-7713-3. DOI: 10.1109/MMSP.2002.1203296.

[2] NGUYEN, N. M., T. T. NGUYEN, H. H. NGUYEN, P.-N. TRAN, and D. N. M. DANG. Voice-Based Age and Gender Recognition: A Comparative Study of LSTM, RezoNet, and Hybrid CNNs-BiLSTM Architecture. *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. 2024, pp. 191–196. ISBN 979-8-3503-6463-7. DOI: 10.1109/ICTC62082.2024.10827387.

[3] MOKGONYANE, T. B., SEFARA, T. J., MODIPA, T. I., MOGALE, M. M., MAN-

AMELA, M. J., and MANAMELA, P. J. Automatic speaker recognition system based on machine learning algorithms. *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*. 2019, pp. 141–146. ISBN 978-1-7281-0369-3. DOI: 10.1109/RoboMech.2019.8704837.

[4] LIN, S., DUAN, R., and MU, X. Design of virtual assistant and human machine dialogue system based on natural language processing algorithms. *2024 International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE)*. 2024, pp. 33–37. ISSN 979-8-3503-9563-1. DOI: 10.1109/EDPEE61724.2024.00012.

[5] YADAV, S.P., GUPTA,A., DOSSANTOSNASCIMENTO,C., DEALBUQUERQUE, V. H. C., NARUKA, M.S., and SINGHCHAUHAN, S. Voice-based virtual-controlled intelligent personal assistants. *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. 2023, pp. 563–568. ISBN 979-8-3503-3802-7. DOI: 10.1109/CICTN57981.2023.10141447.

[6] LIU, Z., FENG, Y., and CHEN, Z. Dialtest: automated testing for recurrent-neural-network-driven dialogue systems. *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 2021, p. 115–126. ISBN 978-1-4503-8459-9. DOI: 10.1145/3460319.3464829.

[7] HASHEM, A., ARIF, M., and ALGHAMDI, M. Speech emotion recognition approaches: A systematic review. *Speech Communication*. 2023, vol. 154, pp. 102974. ISSN 0167-6393. DOI: 10.1016/j.specom.2023.102974.

[8] MA, F., YUAN, Y., XIE, Y., REN, H., LIU, I., HE, Y., REN, F., YU, F. R., and NI, S. Generative technology for human emotion recognition: A scoping review. *Information Fusion*. 2025, vol. 115, pp. 1566–2535. ISSN 0167-6393. DOI: 10.1016/j.inffus.2024.102753.

[9] GEORGE, S.M., and MUHAMEDILYAS, P. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*. 2024, vol. 568, pp. 127015. ISSN 0925-2312. DOI: 10.1016/j.neucom.2023.127015.

[10] ZHENG, X., DU, Y., and QIN, X. Comasa: context multi-aware self-attention for emotional response generation. *Neurocomputing*. 2025, vol. 611, pp. 128692. ISSN 0925-2312. DOI: 10.1016/j.neucom.2024.128692.

[11] MA, Y., ZHANG, Y., BACHINSKI, M., and FJELD, M. Emotion-aware voice assistants: Design, implementation, and preliminary insights. *Proceedings of the Eleventh International Symposium of Chinese*. 2024, p. 527–532. ISBN 979-8-4007-1645-4. DOI: 10.1145/3629606.3629665.

[12] OTHMANI, A., KADOCH, D., BENTOUNES, K., REJAIBI, E., ALFRED, R., and HADID, A. Towards robust deep neural networks for affect and depression recognition from speech. *Pattern Recognition. ICPR International Workshops and Challenges*. 2021, pp. 5–19. ISBN 978-3-0306-8790-8. DOI: 10.1007/978-3-030-68790-8_1.

[13] REJAIBI, E., KOMATY, A., MERIAUDEAU, F., AGREBI, S., and OTHMANI, A. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*. 2022, vol. 71, pp. 103107. ISSN 1746-8094. DOI: 10.1016/j.bspc.2021.103107.

[14] ELSAYED, N., ELSAYED, Z., ASADIZANJANI, N., OZER, M., ABDELGAWAD, A., and BAYOUMI, M. Speech emotion recognition using supervised deep recurrent system for mental health monitoring. *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*. 2022, pp. 1–6. ISBN 978-1-6654-9153-2. DOI: 10.1109/WF-IoT54382.2022.10152117.

[15] GUO, R., GUO, H., WANG, L., CHEN, M., YANG, D., and LI, B. Development and application of emotion recognition technology—a systematic literature review. *BMC Psychology*. 2024, vol. 12, no. 1, pp. 95. ISSN 2050-7283. DOI: 10.1186/s40359-024-01581-4.

[16] LEE, C.-C., MOWER, E., BUSSO, C., LEE, S., and NARAYANAN, S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*. 2011, vol. 53, no. 9, pp. 1162–1171. ISSN 0167-6393. DOI: 10.1016/j.specom.2011.06.004.

[17] LI, L., ZHAO, Y., JIANG, D., ZHANG, Y., WANG, F., GONZALEZ, I., VALENTIN, E., and SAHLI, H. Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 312–317. ISBN 978-0-7695-5048-0. DOI: 10.1109/ACII.2013.58.

[18] BATBAATAR, E., LI, M., and RYU, K. H. Semantic-emotion neural network for emotion recognition from text. *IEEE Access*. 2019, vol. 7, pp. 111866–111878. ISSN 2169-3536. DOI: 10.1109/ACCESS.2019.2934529.

[19] WANI, T. M., GUNAWAN, T. S., QADRI, S. A. A., KARTIWI, M., and AMBIKAIRAJAH, E. A comprehensive review of speech emotion recognition systems. *IEEE Access*. 2021, vol. 9, pp. 47795–47814. ISSN 2169-3536. DOI: 10.1109/ACCESS.2021.3068045.

[20] KHAN, M., GUEAIEB,W., ELSADDIK, A., and KWON, S. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*. 2024, vol. 245, p. 122946. ISSN 0957-4174. DOI: 10.1016/j.eswa.2023.122946.

[21] GHOSH, S., TYAGI, U., RAMANESWARAN, S., SRIVASTAVA, H., and MANOCHA, D. MMER: Multimodal multi-task learning for speech emotion recognition. *Proc. Interspeech*. 2023, p. 1209–1213. DOI: 10.21437/Interspeech.2023-2271.

[22] HE, P., YU, J., GE, C., YU, Y., XU, W., WANG, L., LIU, T., and KAN, Z. Domain-separated bottleneck attention fusion framework for multimodal emotion recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 2025 vol. 21, no. 4. ISSN 1551-6857. DOI: 10.1145/3711865.

[23] LIU, R., ZUO, H., LIAN, Z., SCHULLER, B. W., and LI, H. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*. 2024 vol. 15, no. 4, pp. 1856–1873. ISSN 1949-3045. DOI: 10.1109/TAFFC.2024.3378570.

[24] LI, F., LUO, J., and LIU, W. Speech emotion recognition using multi-modal feature fusion network. *2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. 2023, pp. 884–888. ISBN 979-8-3503-2548-5. DOI: 10.1109/PRAI59366.2023.10332053.

[25] NGUYEN, C.-V. T., MAI, A.-T., LE, T.-S., KIEU, H.-D., and LE, D.-T. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 15154–15167. ISBN 979-8-89176-061-5. DOI: 10.18653/v1/2023.emnlp-main.937.

[26] CHEN, F., SHAO, J., ZHU, S., and SHEN, H. T. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 10761–10770. ISSN 1063-6919. DOI: 10.1109/CVPR52729.2023.01036.

[27] FAN, C., LIN, J., MAO, R., and CAMBIA, E. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*. 2024, vol. 106, p. 102306. ISSN 1566-2535. DOI: 10.1016/j.inffus.2024.102306.

[28] NGUYEN, C.-V. T., KIEU, H.-D., HA, Q.-T., PHAN, X.-H., and LE, D.-T. MI-CGA: Cross-modal graph attention network for robust emotion recognition in the presence of incomplete modalities. *Neurocomputing*. 2025, vol. 623, p. 129342. ISSN 0925-2312. DOI: 10.1016/j.neucom.2025.129342.

[29] PRISAYAD, D., FERNANDO, T., SRIDHARAN, S., DENMAN, S., and FOOKES, C. Dual memory fusion for multimodal speech emotion recognition. in *Interspeech 2023*. 2023, pp. 4543–4547. ISSN 2958-1796. DOI: 10.21437/Interspeech.2023-1090.

[30] KYUNG, J., HEO, S., and CHANG, J.-H. Enhancing multimodal emotion recognition through ASR error compensation and LLM fine-tuning. in *Interspeech 2024*. 2024, pp. 4683–4687. ISSN 2958-1796. DOI: 10.21437/Interspeech.2024-2364.

[31] KHAN, M., TRAN, P.-N., PHAM, N.T., ELSADDIK, A., and OTHMANI, A. MEMOCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*. 2025, vol. 15, no. 1, p. 5473. ISSN 2045-2322. DOI: 10.1038/s41598-025-89202-x.

[32] BAEVSKI, A., ZHOU, Y., MOHAMED, A., and AULI, M. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. in *Advances in Neural Information Processing Systems*. 2020, vol. 33, pp. 12449–12460. ISBN 978-1-7138-2954-6. DOI: 10.48550/arXiv.2006.11477.

[33] ROLLAND, T., and ABAD, A. Introduction to partial fine-tuning: A comprehensive evaluation of end-to-end children's automatic speech recognition adaptation. *Interspeech 2024*. 2024, pp. 5178–5182. ISSN 2958-1796. DOI: 10.21437/Interspeech.2024-1102.

[34] KENTON, J. D. M.-W. C., and TOUTANOVA, L. K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. 2019, pp. 1-16. ISBN 978-1-950737-13-0. DOI: 10.18653/v1/N19-1423.

[35] FAN, W., WANG, X., and WU, Y. Diversified top-k graph pattern matching. *Proceedings of the VLDB Endowment (PVLDB)*. 2013, vol. 6, no. 13, pp. 1510–1521. ISSN 2150-8097. DOI: 10.14778/2536258.2536263.

[36] VELIČKOVI'C, P., CUCURULL, G., CASANOVA, A., ROMERO, A., LI'O, P., and BENGIO, Y. Graph attention networks. *International Conference on Learning Representations*. 2018. DOI: 10.48550/arXiv.1710.10903.

[37] ZHANG, S., TONG, H., XU, J., and MACIEJEWSKI, R. Graph convolutional networks: A comprehensive review. *Computational Social Networks*. 2019, vol. 6, no. 1, pp. 1–23. ISSN 2197-4314. DOI: 10.1186/s40649-019-0069-y.

[38] BUSSO, C., BULUT, M., LEE, C.-C., KAZEMZADEH, A., MOWER, E., KIM, S., CHANG, J. N., LEE, S., and NARAYANAN, S. S. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*. 2008, vol. 42, pp. 335–359. ISSN 1574-0218. DOI: 10.1007/s10579-008-9076-6.

[39] PORIA, S., HAZARIKA, D., MAJUMDER, N., NAIK, G., CAMBIA, E., and MIHALCEA, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 527–536. ISBN 978-15-108-9099-2. DOI: 10.18653/v1/P19-1050.

[40] ZHANG, Z., and SABUNCU, M. Generalized cross-entropy loss for training deep neural networks with noisy labels. in *Advances in Neural Information Processing Systems*. 2018, vol. 31.

[41] SANTOSO, J., YAMADA, T., ISHIZUKA, K., HASHIMOTO, T., and MAKINO, S. Speech emotion recognition based on self-attention weight correction for acoustic and text features. *IEEE Access*. 2022, vol. 10, pp. 115732–115743. ISSN 2169-3536. DOI: 10.1109/ACCESS.2022.3219094.

[42] ZHAO, J., LI, R., JIN, Q., WANG, X., and LI, H. MEMOBERT: Pre-training model with prompt-based learning for multimodal emotion recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 4703–4707. ISSN 2379-190X. DOI: 10.1109/ICASSP43922.2022.9746910.

[43] NGUYEN, L. H., PHAM, N. T., KHAN, M., OTHMANI,A., and ELSADDIK, A. HUBERT-CLAP: Contrastive learning-based multimodal emotion recognition using self-alignment approach. in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*. 2024, pp.1–6. ISBN 979-84-0071-273-9. DOI: 10.1145/3696409.3700183.

[44] MA, Z., ZHENG, Z., YE, J., LI, J., GAO, Z., ZHANG, S., and CHEN, X. Emotion2Vec: Self-supervised pre-training for speech emotion representation. in *Findings of the Association for Computational Linguistics: ACL 2024*. 2024, pp. 15747–15760. ISBN 979-83-3130-1828. DOI: 10.18653/v1/2024.findings-acl.931.

[45] QI, X., WEN, Y., ZHANG, P., and HUANG, H. MFGCN: Multimodal fusion graph convolutional network for speech emotion recognition. in *Neurocomputing*. 2025, vol. 611, p. 128646. ISSN 0925-2312. DOI: 10.1016/j.neucom.2024.128646.

[46] LI, J., WANG, X., LV, G., and ZENG, Z. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*. 2023 vol. 550, p. 126427. ISSN 0925-2312. DOI: 10.1016/j.neucom.2023.126427.

[47] LI, J., WANG, X., LV, G., and ZENG, Z. GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*. 2024, vol. 26, pp. 77–89. ISSN 1520-9210. DOI: 10.1109/TMM.2023.3260635.

[48] LU, N., HAN, Z., HAN, M., and QIAN, J. Bi-stream graph learning based multimodal fusion for emotion recognition in conversation. *Information Fusion*. 2024, vol. 106, p. 102272. ISSN 1566-2535. DOI: 10.1016/j.inffus.2024.102272.

[49] WU, F., SOUZA, A., ZHANG, T., FIFTY, C., YU, T., and WEINBERGER, K. Simplifying graph convolutional networks. *International Conference on Machine Learning*. 2019, pp. 6861–6871. DOI: 10.48550/arXiv.1609.02907.

[50] HAMILTON, W., YING, Z., and LESKOVEC, J. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*. 2017, vol. 30. DOI: 10.48550/arXiv.1706.02216.