EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems

Research Article EAI.EU

HyperDyG: Hypergraph-Driven Dynamic Fusion for Semi-Supervised Multimodal Emotion Recognition

Nhut Minh Nguyen[1], Trung Thanh Nguyen[1], Thu Thuy Le[1], Ngoc-Hanh Dang[2,4], Phuong Luu Vo[3,4], Lam Thanh Hien[5], Duc Ngoc Minh Dang[1,*]

[1] AiTA Lab, Faculty of Information Technology, FPT University, Ho Chi Minh City, Vietnam
[2] Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[3] School of Computer Science and Engineering, International University, Ho Chi Minh City, Vietnam
[4] Vietnam National University, Ho Chi Minh City (VNU-HCM), Vietnam
[5] Faculty of Information Technology, Lac Hong University, Dong Nai Province, Vietnam

Abstract

Speech emotion recognition (SER) is important in healthcare, education, human–computer interaction, and customer service. Multimodal emotion recognition (MER) integrates audio and textual modalities to achieve a comprehensive understanding of human affect, but still suffers from limited labeled data and complex cross-modal relations. To address these challenges, we propose HyperDyG, a dynamic hypergraph-driven MER framework. The HyperDyG leverages the strengths of dynamic hypergraph learning (DHL), cross-modal transformer (CMT), and an adaptive gated multimodal unit (GMU) for robust multimodal fusion. HyperDyG is further enhanced with a semi-supervised learning strategy that incorporates weak–strong augmentation, confidence-filtered pseudo-labeling, and consistency regularization to effectively exploit large-scale unlabeled data. The HyperDyG achieves state-of-the-art (SOTA) performance on the benchmark emotion dataset and maintains stable accuracy across varying unlabeled ratios. The findings of HyperDyG highlight the effectiveness and scalability of the proposed architecture in real-world low-label MER scenarios.

Received on 13 November 2025; accepted on 10 February 2026; published on 19 February

Keywords: Hypergraph learning, Multimodal fusion, Cross-modal transformer, Dynamic gating, Semi-supervised emotion recognition

doi:10.4108/eetinis.131.10903

1. Introduction

Speech emotion recognition (SER) is a crucial task in the field of Human-Computer Interaction (HCI), where computer systems can detect emotional states through signal processing [1]. SER models utilize features from the voice, including pitch, energy, rhythm, intensity variation, and time–frequency spectral characteristics [2]. SER technology plays an essential role in healthcare, smart education, customer service, intelligent conversation technology, and psychological support systems. However, SER uses only the acoustic modality, making the model effective in the presence of environmental noise, speaker variability, and acoustic distortions, which significantly affect emotional cues. Additionally, speech-only emotion recognition often lacks contextual information and suffers from inherent ambiguity in emotional expression, as a single utterance may be insufficient to capture the semantic nuances or intensity of sentiment [3, 4]. Therefore, multimodal emotion recognition (MER) has emerged to address these challenges by incorporating additional modalities to enhance emotional understanding and improve robustness.

MER integrates multiple types of information, including audio, text, visual signals, and electroencephalogram data, to facilitate comprehensive learning of emotional states [5, 6]. MER reduces prediction errors, enhances model generalization, and captures deeper cross-modal relationships by coordinating multiple information

[*] Corresponding author. Email: ducdnm2@fe.edu.vn

sources [7]. In MER, several novel techniques are applied, including cross-modal attention for semantic alignment between text and audio [8, 9], contrastive learning to minimize modality-specific representation gaps [10–12], and graph-based modeling to capture the relational structures among multimodal features [13–15]. However, the limitation of MER is the modality fusion. Traditional fusion strategies rely on static mathematical operations. These operations fail to capture the dynamic and contextual contribution of each modality. The overall robustness of the model is reduced, particularly when certain modalities are noisy, overly dominant, or semantically ambiguous.

Additionally, in real-world conditions, most MER data remain unlabeled due to the high cost and subjectivity of manual annotation. The scarcity of labeled samples leads to overfitting, limits the model's ability to learn cross-modal dependencies, and reduces generalization [16]. To address these challenges, semi-supervised learning is introduced as an effective paradigm that leverages both labeled and unlabeled data [17, 18]. The semi-supervised frameworks such as Pseudo-Labeling [19], Mean Teacher [20], and Noisy Student [21] emphasize prediction stability and teacher–student consistency. The FixMatch [22] is a novel framework for semi-supervised learning that combines confidence-based pseudo-labeling with weak-strong data augmentation. These approaches generate pseudo-labels, enforce consistency, and introduce augmentation diversity, enabling models to learn from noisy samples, reduce representation drift, and expand the decision boundary in latent space. When integrated with cross-modal modeling, semi-supervised learning further stabilizes MER and significantly enhances effectiveness in scenarios with low-label or heterogeneous data [23, 24].

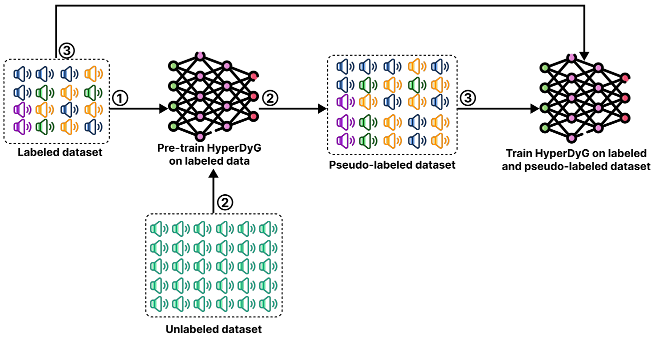

③
Labeled dataset
①
Pre-train HyperDyG on labeled data
②
Pseudo-labeled dataset
③
Train HyperDyG on labeled and pseudo-labeled dataset
②
Unlabeled dataset

Figure 1. The overview workflow of HyperDyG architecture

To address these challenges, we propose HyperDyG, a dynamic hypergraph-based MER architecture that integrates a cross-modal transformer and a gated multimodal fusion mechanism. In this architecture, we employ two modalities, text and audio, as the primary sources for emotion prediction. HyperDyG captures higher-order structural dependencies between nodes through Dynamic Hypergraph Learning (DHL), enabling the model to learn multidimensional relational patterns that are often overlooked in traditional pairwise graph representations. In parallel, the Cross-Modal Transformer (CMT) performs bidirectional alignment across modalities, enhancing contextual relevance and cross-modal coherence. The Gated Multimodal Unit (GMU) adaptively regulates the contribution of each modality based on contextual cues, thereby overcoming the limitations of static fusion methods. Furthermore, we apply a semi-supervised approach on the HyperDyG architecture with 3 stages, visualized in Figure 1. First, the model is pre-trained on the labeled dataset. Second, unlabeled data are passed through the model to generate high-confidence pseudo-labels using a pseudo-threshold. Third, the labeled and pseudo-labeled samples are combined to retrain HyperDyG for improved performance.

In this paper, our contributions are as follows:

- Introducing a novel MER architecture, HyperDyG, that leverages audio and text modalities and integrates a dual-stream design combining DHL and CMT.
- Designing a GMU-based fusion strategy that adaptively balances structure-aware hypergraph representations and context-aware cross-attention features to handle varying modality reliability.
- Developing a semi-supervised learning framework for MER by incorporating weak–strong augmentation, confidence-filtered pseudo-labeling, and consistency regularization to better leverage unlabeled data.
- Achieving state-of-the-art (SOTA) performance for MER in supervised settings on two benchmark datasets, Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Emotional Speech Dataset (ESD).

The structure of this paper is organized as follows: Section 2 reviews related work on MER and the underlying technologies. Section 3 presents the proposed HyperDyG architecture and its key components. Section 4 describes the experimental setup, datasets, evaluation metrics, and implementation details. Section 5 reports and discusses the experimental results, including ablation studies and comparative analyses. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related works

2.1. Graph-based MER approaches

Nguyen *et al.* [13] proposed a dual-stream framework that combined a Cross-modal Heterogeneous Graph Attention Network (CH-GAT) with a Cross-modal Convolutional Block Attention Mechanism (xCBAM). CH-GAT modeled intra- and inter-modal dependencies through heterogeneous graph reasoning, while xCBAM refined emotionally salient cues via cross-modal channel and spatial attention. This combination enabled more discriminative feature representations and enhanced performance on benchmark datasets.

Qi *et al.* [14] proposed a multimodal fusion graph convolutional network that modeled SER as message passing on a multimodal graph. The model constructed acoustic and textual word nodes and defined intra-modal edges that encoded sentiment, semantic, and temporal dependencies, allowing graph convolution to capture intrinsic structure within each modality. The mechanism introduced inter-modal acoustic–textual edges together with a multi-perspective fusion module that retained supplementary unimodal cues while propagating information across modalities. A multi-angle loss was jointly optimized for multimodal and unimodal branches, so that each modality-specific graph remained discriminative and the fused representation was strengthened.

Fan *et al.* [15] presented a fusion framework that employed a dual-channel Bidirectional Long Short-Term Memory to capture temporal features from modality pairs. These features were represented as graphs using a graph convolutional network with speaker embeddings. A novel density loss reduced redundant information, yielding more distinct and comprehensive representations. Evaluations on benchmark datasets confirmed superior performance over existing SOTA baselines.

Although these methods effectively exploit graph structures, they are still primarily built on pairwise graphs, where each edge links only two nodes. This design limits their ability to capture higher-order interactions. Pairwise graphs often rely on manually designed edge types or fixed similarity thresholds and are typically static. When hypergraphs are constructed dynamically, they can adapt to the connectivity of each utterance and better reflect the evolving audio–text interactions. These properties make dynamic hypergraph modeling a compelling choice for MER, motivating the design adopted in our work.

2.2. Feature fusion in MER approaches

Khan *et al.* [8] presented MemoCMT, a cross-modal transformer that integrated audio features from HuBERT and textual features from BERT. The architecture explored multiple aggregation techniques (CLS, mean, max, and min), with MIN yielding the best results. Evaluations on benchmark datasets showed that MemoCMT achieved SOTA weighted and unweighted accuracies, advancing robust MER.

Kyung *et al.* [25] introduced a robust MER system to handle Automatic Speech Recognition (ASR) induced textual errors. The approach combined an ASR error compensation strategy with preference learning-based fine-tuning of a large language model (LLM). A cross-modal transformer fused speech and ASR-generated text embeddings, while the Kullback–Leibler divergence aligned ASR text with the ground truth. RankNet-based preference learning further improved the LLM's sensitivity to emotional nuances. Experiments on the emotional dataset confirmed significant gains, especially under high ASR error rates.

Wang *et al.* [26] proposed a modality-sensitive MER framework that explores complementary features from pre-trained acoustic and linguistic representations. They employ pre-trained transformer-based encoders to obtain self-supervised audio and text embeddings, followed by modality-specific transformer encoders and self-attention–based frame weighting to emphasize emotionally salient segments. A self-attention–based fusion mechanism, together with a modality interaction transformer equipped with learnable emotion query tokens, is then used to integrate the two modalities and construct utterance-level representations.

Despite these advances, most multimodal fusion strategies still rely on statistical fusion. Modalities are combined using fixed mathematical operators such as concatenation, summation, element-wise product, or simple pooling over pre-aligned features. These static schemes implicitly assume that all modalities contribute equally to the overall outcome. They fail to adapt when one modality becomes noisy, missing, or semantically ambiguous. Attention-based fusion enhances flexibility by learning importance weights for individual tokens or frames. However, most existing designs operate only at the feature level. They do not explicitly model higher-order interactions among modalities or provide fine-grained control over modality dominance in MER. These limitations motivate adaptive fusion architectures that can dynamically regulate modality contributions and exploit both contextual and structural dependencies across modalities.

2.3. Semi-supervised MER approaches

Zhang *et al.* [16] proposed a framework that combined cross-modal knowledge transfer and semi-supervised learning to address the lack of high-quality labeled data in SER. The method exploited large amounts of unlabeled audio-visual data. Emotion cues from facial expressions were transferred to the audio modality.

At the same time, pseudo-labels were generated from the audio stream using a FixMatch-based semi-supervised approach. Two fusion strategies were designed. Consistent and random retention of consistent cross-modal labels and random inclusion of additional visual labels. Weighted fusion aggregated soft label distributions from both modalities.

Tsouvalas *et al.* [27] introduced the first privacy-preserving SER framework based on semi-supervised federated learning. Unlike centralized SER systems, their method kept raw speech data on user devices and trained local models collaboratively via the Federated averaging algorithm. To address the scarcity of labeled data, they integrated a self-training mechanism that generated pseudo-labels from unlabeled on-device speech and combined them with limited labeled samples. The model further employed a spectro-temporal channel attention mechanism to capture emotional cues without increasing computational complexity. This work highlighted the feasibility of federated SER, even under highly non-IID data distributions, while ensuring data privacy.

Agarla *et al.* [23] addressed the challenge of cross-lingual SER, where models trained on one language typically degraded when applied to another. They proposed a semi-supervised learning framework that combined labeled data from a source language with large amounts of unlabeled data from a target language. The method relied on pseudo-labeling, experimenting with both high-confidence predicted labels and distribution-based strategies, and further introduced an utterance rebalancing mechanism to mitigate the imbalance between source and target corpora.

These approaches emphasize the significance of semi-supervised learning for MER. Multiple modalities must be annotated and aligned simultaneously. This process often occurs under severe class imbalance and distribution shifts across domains. Therefore, leveraging abundant unlabeled multimodal speech while maintaining robust fusion is crucial for building scalable and generalizable MER systems. This motivates the development of semi-supervised MER architectures that can jointly exploit labeled and unlabeled data, adapt to modality-specific noise, and preserve discriminative cross-modal representations.

3. Methodology

The overall HyperDyG framework is illustrated in Figure 2. In stage A, the supervised HyperDyG architecture first encodes audio and text using pre-trained WavLM and BERT models, respectively. The resulting embeddings are fed into two parallel streams: the DHL module constructs dual heterogeneous hypergraphs to capture higher-order intra- and inter-modality dependencies, while the CMT module performs bidirectional cross-modal attention to model fine-grained contextual alignment. Their outputs are fused by the GMU, which adaptively regulates modality contributions and produces a joint representation for emotion classification. Stage B illustrates the semi-supervised pipeline, where weakly and strongly augmented views of unlabeled audio–text pairs are processed through HyperDyG to obtain predictions. Stage C illustrates pseudo-labeling and consistency regularization, where high-confidence predictions on weakly augmented samples are used as pseudo-labels to supervise their strongly augmented counterparts, which are then optimized jointly with the labeled data.

3.1. Modality embedding extraction

For the acoustic representation, we employ the WavLM [28] for the audio encoder. The transformer-based self-supervised model is pre-trained on large-scale multilingual corpora. WavLM effectively captures both low-level spectral cues, such as pitch, energy, or timbre, and high-level semantic patterns relevant to emotional expression. Each input utterance is converted into a fixed-length embedding vector by averaging the hidden states from the last transformer layer. For textual representation, we adopt BERT [29] to generate rich context-aware embeddings by modeling token-level dependencies. This allows the model to capture fine-grained linguistic emotion cues, including sentiment polarity, semantic intensity, and nuanced contextual variations within utterances.

To adapt the pre-trained encoders to emotion recognition, we apply partial fine-tuning. Only the top transformer layers of WavLM and BERT are fine-tuned during multimodal training, whereas the lower layers are frozen to preserve the generic acoustic and linguistic knowledge learned from large-scale corpora. This strategy ensures efficient optimization and enhances emotion-specific sensitivity in the embeddings.

After extracting the features, the feature embeddings of text and audio are presented in Eq. (1), which serve as the input representations for subsequent CMT and DHL modules.

$$F_t = \{u_1^t, u_2^t, \ldots, u_N^t\}, \quad F_a = \{u_1^a, u_2^a, \ldots, u_N^a\}, \tag{1}$$

where u_i^t and u_i^a denote the utterance-level embeddings for the i-th utterance in the text and audio modalities, respectively, and N is the number of utterances.

3.2. Dynamic hypergraph learning (DHL)

In the proposed HyperDyG architecture, the DHL module operates in parallel with the CMT module. The DHL module performs structural-level modeling by constructing and learning a dynamic hypergraph [30] to capture higher-order dependencies both within and across modalities. This parallel design enables

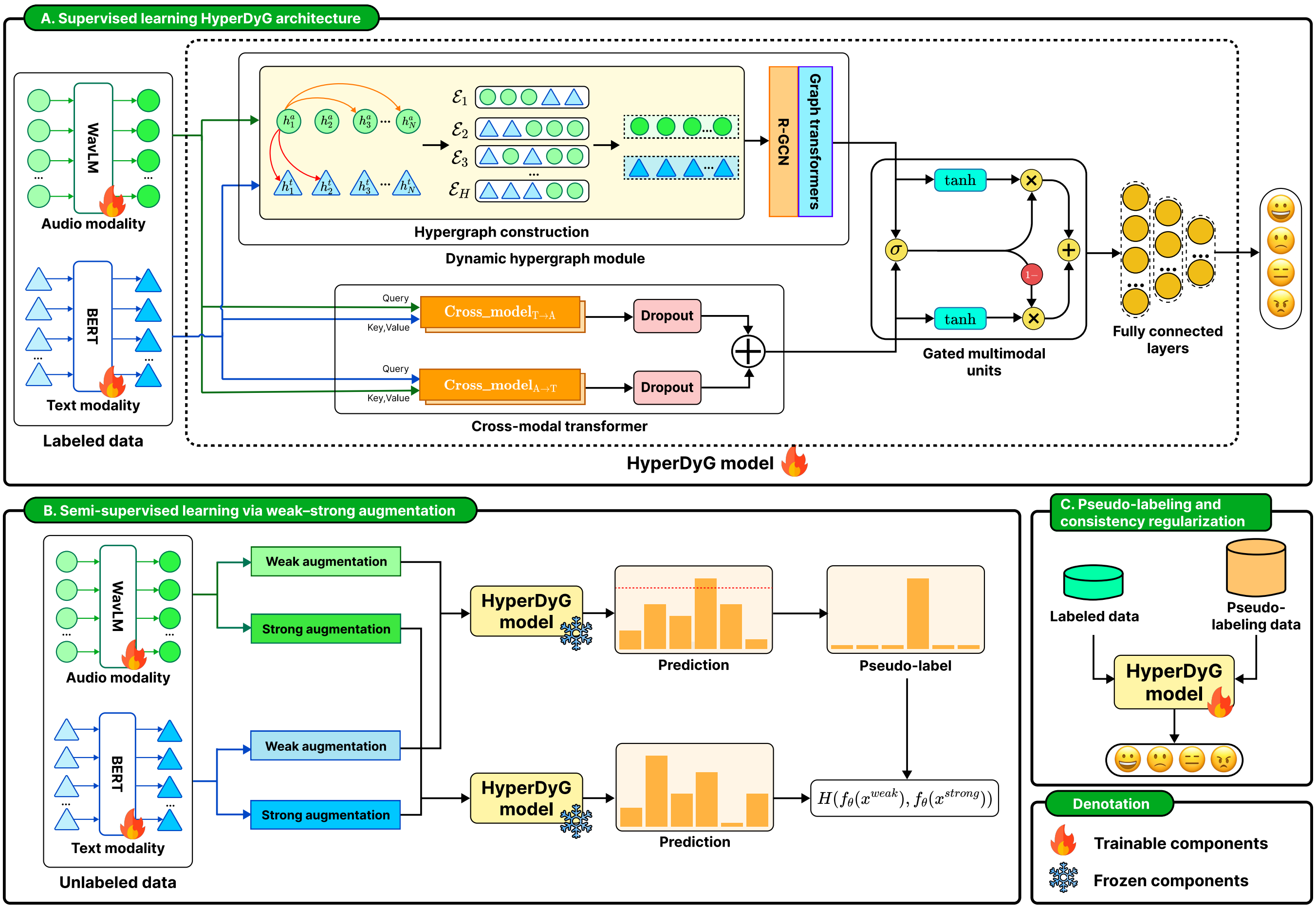


Figure 2. Detailed overview of the HyperDyG framework, showing the supervised pipeline with a dual-stream architecture combining DHL and CMT, together with the semi-supervised process where augmented unlabeled samples are processed to generate predictions for pseudo-labelin

HyperDyG to jointly benefit from fine-grained feature exchange and relational reasoning, leading to a more comprehensive multimodal representation. The DHL process consists of two main components: hypergraph construction and graph learning.

Hypergraph construction. To overcome the limitation of pairwise graph structures for modeling simple binary relations between nodes, hypergraphs allow each hyperedge to connect multiple nodes simultaneously. The hypergraph construction enables the model to encode higher-order relationships, thereby enhancing multimodal representation learning and contextual integration across modalities. Given the audio and text feature embeddings F_a and F_t, the DHL module dynamically constructs a dual heterogeneous hypergraph to model both intra- and inter-modal dependencies. Each modality is first represented as a node set, which is presented in Eq.(2).

$$\mathcal{V}_t = \{u_1^t, u_2^t, \dots, u_N^t\}, \quad \mathcal{V}_a = \{u_1^a, u_2^a, \dots, u_N^a\}, \tag{2}$$

where each node corresponds to an embedding vector in either the text or audio modality.

For each node, the Top-K most similar nodes from the opposite modality are selected to form heterogeneous cross-modal hyperedges, enabling multi-node aggregation between semantically and acoustically correlated instances. In contrast, the homogeneous connections capture modality-internal correlations, such as semantic coherence within text and acoustic smoothness within audio, as expressed in Eq. (3).

$$\mathcal{E}_t^{intra} = \frac{1}{N}\sum_{i=1}^{N_t} u_i^t, \quad \mathcal{E}_a^{intra} = \frac{1}{N}\sum_{i=1}^{N_a} u_i^a. \tag{3}$$

To capture inter-modal relationships, similarity between the normalized node embeddings of the two modalities is computed through dot-product attention. The inter-relationship is formulated in Eq. (4).

$$\mathbf{S}_{t\to a} = F_t F_a^\top, \quad \mathbf{S}_{a\to t} = F_a F_t^\top. \tag{4}$$

For each node, the Top-K most similar nodes from the opposite modality are selected to form heterogeneous cross-modal hyperedges, enabling joint aggregation between semantically or acoustically correlated instances. In Eqs. (5) and (6), the inter-modal hyperedge construction with mean pooling is used as a lightweight aggregator. Mean pooling summarizes the features of the Top-K algorithm of the cross-modal neighbors into a single hyperedge representation. The higher-order relations are primarily induced by the similarity-based neighborhood selection, which determines which nodes are grouped into each hyperedge.

$$\mathcal{E}_t^{inter} = \text{Mean}\left([u_i^t, \text{Top-K}_{t\to a}(u_j^a)]\right), \tag{5}$$

$$\mathcal{E}_a^{inter} = \text{Mean}\left([u_j^a, \text{Top-K}_{a\to t}(u_i^t)]\right). \tag{6}$$

Finally, the intra- and inter-modal hyperedges are fused to form dual heterogeneous hyperedge representations that jointly capture both local modality-specific context and cross-modality associations. The feature representation for heterogeneous graph construction is defined in Eqs. (7) and (8).

$$H_t^{hyper} = \frac{1}{2}(\mathcal{E}_t^{intra} + \mathcal{E}_t^{inter}), \tag{7}$$

$$H_a^{hyper} = \frac{1}{2}(\mathcal{E}_a^{intra} + \mathcal{E}_a^{inter}). \tag{8}$$

The graph construction enables the model to dynamically adapt hyperedge connections based on the evolving similarity between modalities, thereby laying the foundation for subsequent hypergraph learning and multimodal fusion in HyperDyG.

After hyperedge construction, the DHL module performs message passing to propagate information between nodes and hyperedges in a node–edge–node manner. This mechanism enables the model to capture both local and global relational structures across modalities, complementing the direct attention-based interactions of the CMT module. The propagation process consists of two main stages:

- In the node-to-edge phase, node embeddings are aggregated to form hyperedge representations, allowing the model to learn higher-order contextual patterns by combining information from multiple semantically or acoustically related nodes. This stage effectively captures group-level correlations, such as emotion-consistent clusters across utterances or acoustic segments.

- In the edge-to-node phase, each node updates its representation by averaging the features of the hyperedges it participates in, enabling it to absorb both intra-modal coherence and inter-modal associations propagated through the hypergraph structure. In our implementation, we apply mean pooling to incident nodes for computational efficiency and robustness.

This update enriches each node's embedding with cross-modal contextual cues, enhancing emotional alignment between modalities. Through this propagation, the DHL process facilitates hierarchical relational reasoning, effectively bridging the structural gap between modalities and reinforcing the synergy between attention-based and structure-based learning that underpins the HyperDyG architecture.

Graph learning. The graph learning module refines node representations through relational modeling, utilizing the hypergraph construction via Relational Graph Convolutional Networks (R-GCN) [31] and Graph Transformer [32] layers. This stage enables HyperDyG to capture complementary relational patterns, localized dependencies through graph convolution, and global contextual interactions via transformer-based attention, thereby enhancing the multimodal representations before fusion.

The R-GCN layer performs the neighborhood aggregation across the dual heterogeneous hypergraph, where each relation type is associated with a distinct transformation matrix. For the node u_i^m belonging to modality m, the representation of this node is presented in Eq (9).

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right), \tag{9}$$

where $\mathcal{R}$ denotes the set of relation types, $\mathcal{N}_i^r$ is the neighborhood of node i under relation r, $c_{i,r}$ is a normalization term, and $W_r^{(l)}$ and $W_0^{(l)}$ are learnable weight matrices. In HyperDyG, there are three relation types to guide R-GCN aggregation. The identity relations help preserve the self-connection of each node. The intra-modal relations capture the node dependencies within the same modality, and the inter-modal relations model cross-modal interactions. This design enables R-GCN to effectively integrate both structural and cross-modal relational cues.

The feature representation of R-GCN is subsequently passed to a Graph Transformer layer, which captures long-range dependencies and global structural interactions that extend beyond immediate neighborhoods. The Graph Transformer applies self-attention over the graph topology. Unlike standard attention that treats all tokens equally, the graph-based attention preserves topological constraints, allowing the model to focus on structurally and semantically relevant nodes. After the graph transformer, the final feature representation produced by the DHL module is presented in Eq. (10).

$$F_{DHL} = \text{Transformer}(\text{R-GCN}(Z)), \tag{10}$$

where Z denotes the relation-aware embeddings obtained from the R-GCN layers.

With the R-GCN and Graph Transformer layers, the graph learning module jointly captures relation-aware local dependencies and global contextual structures. The resulting node embeddings encode rich multimodal relationships that provide a strong foundation for the GMU to perform adaptive fusion.

3.3. Cross-modal transformer (CMT)

The CMT module aims to enable fine-grained interaction between the audio and text modalities before integrating them into the hypergraph construction stage. Parallel to the DHL module, this module focuses on embedding-level alignment by allowing both modalities to attend to one another through bidirectional attention mechanisms.

Given the modality embeddings F_a and F_t, the CMT module performs two complementary cross-attention operations: audio-to-text $(A \rightarrow T)$ and text-to-audio $(T \rightarrow A)$. These operations ensure that each modality incorporates contextual cues from the other, thereby enhancing consistency in emotional representation across modalities.

In the first stage, the audio-to-text attention, which is presented in Eq. (11), treats F_a as the query and F_t as the key-value pair, allowing the speech modality to attend to linguistically relevant text tokens.

$$\mathrm{CM}_{A\rightarrow T} = \mathrm{Softmax}\left(\frac{F_a \mathbf{W}_{Q_a} \left(\mathbf{W}_{K_t}\right)^{\top} (F_t)^{\top}}{\sqrt{d_k}}\right) F_t \mathbf{W}_{V_t}, \quad (11)$$

where $\mathbf{W}_{Q_a}$, $\mathbf{W}_{K_t}$, and $\mathbf{W}_{V_t}$ are trainable projection matrices for query, key, and value spaces, respectively, and d_k is the key dimension used for normalization.

Similarly, text-to-audio attention, defined in Eq. (12), uses F_t as the query and F_a as the key-value pair, enabling the text modality to integrate relevant paralinguistic and prosodic cues from speech.

$$\mathrm{CM}_{T\rightarrow A} = \mathrm{Softmax}\left(\frac{F_t \mathbf{W}_{Q_t} \left(\mathbf{W}_{K_a}\right)^{\top} (F_a)^{\top}}{\sqrt{d_k}}\right) F_a \mathbf{W}_{V_a}, \quad (12)$$

where $\mathbf{W}_{Q_t}$, $\mathbf{W}_{K_a}$, and $\mathbf{W}_{V_a}$ correspond to the text modality's query, key, and value transformations.

After each attention operation, a dropout layer is applied to prevent overfitting and ensure regularization. The features from both directions are then concatenated to form the final multimodal representation, as shown in Eq. (13).

$$F_{CMT} = [\mathrm{CM}_{A\rightarrow T} \,\|\, \mathrm{CM}_{T\rightarrow A}], \quad (13)$$

where $\|$ denotes concatenation. The resulting F_{CMT} serves as a unified embedding containing complementary emotional information from both modalities.

3.4. Gated multimodal fusion

To adaptively combine the structure-aware representation from DHL and the context-aware representation from CMT, we employ a feature-wise GMU. In the GMU workflow, the gating vector z, defined in Eq. (14), determines the relative contribution of each modality during fusion. These functions act as a soft attention controller, assigning adaptive weights to the features derived from the DHL and CMT modules based on their contextual relevance.

$$z = \sigma(W_g[F_{DHL} \| F_{CMT}] + b_g), \quad (14)$$

where $\sigma(\cdot)$ denotes the sigmoid activation, $\|$ represents feature concatenation, and W_g and b_g are learnable parameters. The z is a feature-wise gating vector $z \in \mathbb{R}^d$, $W_g \in \mathbb{R}^{d\times 2d}$. With feature-wise gating, each embedding dimension is adaptively weighted via element-wise gating, rather than being controlled by a single global scalar.

The final fused representation of the GMU is formulated in Eq. (15).

$$F_{fused} = z \odot \tanh(F_{DHL}) + (1 - z) \odot \tanh(F_{CMT}), \quad (15)$$

where $\odot$ denotes element-wise multiplication. The F_{DHL} and F_{CMT} are feature representations obtained from the DHL and CMT modules, respectively. When z is close to 0, the fusion prioritizes information from the CMT branch, whereas values is close to 1 favor representations from the DHL branch. The tanh activation introduces non-linearity, promoting smoother feature interactions and stabilizing the scaling of the fused output.

Through this adaptive gating mechanism, the GMU directly addresses the limitations of traditional fusion strategies by dynamically balancing the contributions of contextual and relational information. By integrating context-aware features from the CMT module with structure-aware representations from the DHL module, the GMU enables flexible, content-dependent fusion that adapts to variations in modality reliability. This design ensures that the model emphasizes the most expressive modality under different emotional contexts, thereby achieving a unified, discriminative, and robust multimodal representation for emotion classification.

3.5. Semi-supervised learning framework

The HyperDyG employs a semi-supervised learning framework that leverages both labeled and unlabeled data, thereby enhancing representation robustness and overall generalization. First, HyperDyG is pre-trained using the labeled dataset to establish a stable multimodal representation space. After that, the unlabeled data are passed through the pre-trained model to obtain high-confidence pseudo-labels, which serve as supervisory signals for reliable samples. Finally, the labeled and pseudo-labeled datasets are combined to retrain HyperDyG, enabling the model to refine decision boundaries and benefit from a substantially larger training set. The framework is divided into 2 parts: supervised pre-training and semi-supervised learning via weak–strong augmentation.

Supervised pre-training. In the first stage, the HyperDyG architecture is trained on labeled data to establish a reliable multimodal representation space. With the labeled dataset $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{N_L}$, the model is optimized using a class-weighted cross-entropy loss that compensates for emotional category imbalance. The function of supervised learning loss is presented in Eq. (16).

$$\mathcal{L}_{CE} = -\frac{1}{N_L} \sum_{i=1}^{N_L} w_{y_i} \log p_\theta(y_i|x_i), \tag{16}$$

where w_{y_i} compensates for emotional class imbalance.

This stage enables the CMT and DHL modules to learn context-aware and structure-aware embeddings, respectively, while the GMU fuses them into a unified, discriminative multimodal representation. This pre-trained model subsequently serves as a strong initialization for effectively exploiting unlabeled data in the semi-supervised learning stage.

Semi-supervised learning via weak–strong augmentation. We employ a FixMatch-inspired strategy [22] in the HyperDyG architecture to leverage real-world emotion corpora, which often contain abundant unlabeled multimodal samples. The framework integrates pseudo-labeling and consistency regularization to effectively utilize these data. In this approach, confident model predictions on weakly augmented inputs are used as pseudo-labels to guide the labeling of unlabeled samples. The model is then trained to produce consistent predictions across weakly and strongly augmented versions, thereby improving stability under perturbations.

We use two augmented views for each unlabeled input: a weak version x^{weak} for stable pseudo-labeling and a strong version x^{strong} for consistency training. The weakly augmented sample is used to generate a reliable pseudo-label, as shown in Eqs. (17) and (18).

$$\hat{y}_j = \arg\max_k \, p_\theta(y_k|x_j^{weak}), \tag{17}$$

$$c_j = \max_k \, p_\theta(y_k|x_j^{weak}), \tag{18}$$

where c_j denotes the prediction confidence. Only samples with confidence above a threshold τ are retained, forming a pseudo-labeled subset $\hat{\mathcal{D}}_U$. These pseudo-labeled samples contribute to a confidence-weighted cross-entropy loss, as defined in Eq. (19).

$$\mathcal{L}_{semi} = \frac{1}{|\tilde{\mathcal{D}}_U|} \sum_{(x_j,\tilde{y}_j,c_j)\in\hat{\mathcal{D}}_U} c_j \, \mathcal{L}_{CE}(x_j^{strong}, \hat{y}_j). \tag{19}$$

Weak and strong augmentations are applied to both text and audio modalities to enhance model robustness. For the text modality, weak augmentation applies low-rate word dropout, while strong augmentation increases the dropout intensity to introduce greater variability. For the audio modality, Gaussian noise is added to the embeddings, with weak augmentation using mild perturbations and strong augmentation applying more substantial distortions. These transformations simulate natural variations in speech and language, improving the model's resilience to noise and contextual fluctuations.

The overall loss function of the semi-supervised HyperDyG framework is presented in Eq. (20), integrating the supervised, pseudo-label, and consistency terms.

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_u \mathcal{L}_{semi}, \tag{20}$$

where λ_u is a balance coefficients th at re gulate the influence of unlabeled data.

The semi-supervised HyperDyG framework encourages the model to expand decision boundaries into high-confidence regions while maintaining prediction stability under perturbations, effectively utilizing both labeled and unlabeled data. By integrating weak–strong augmentation with pseudo-label consistency, HyperDyG achieves robust, context-aware, and structure-consistent learning even in low-label scenarios.

4. Experiment settings

4.1. Datasets

We employ the IEMOCAP and ESD datasets to evaluate the HyperDyG architecture. The dataset varies in terms of language, speaker diversity, and recording conditions. Their key statistics are summarized in Table 1.

Table 1. The statistics of two experimental datasets

Datasets	Number of classes	Utterances			Duration (h)
		Train	Valid	Test	
IEMOCAP	4	4,479	498	554	12
ESD	5	13,996	1,750	1,750	13.44

The IEMOCAP dataset [33], published in 2008, is a multimodal, multi-speaker dataset collected at the SAIL Lab at the University of Southern California. The IEMOCAP dataset is widely used as a benchmark for emotion recognition tasks, which comprises approximately 12 hours of audiovisual data. We divide the dataset into four classes: anger (1,103 utterances), happiness (1,636 utterances), neutral (1,708 utterances), and sadness (1,084 utterances).

The ESD dataset [34] developed by the National University of Singapore and the Singapore University of Technology and Design addresses the growing need for high-quality emotional voice conversion research. ESD comprises 350 parallel utterances from 10 native English and 10 native Mandarin speakers, covering five emotions: neutral, happiness, anger, sadness, and surprise. In this study, we utilize the English subset of the ESD dataset to evaluate our proposed approach for emotional voice conversion.

4.2. Implementation details

The experiments using the HyperDyG architecture are implemented with the PyTorch framework and conducted on an NVIDIA Tesla V100 GPU. To ensure reproducibility, each experiment is repeated 5 times with different random seeds, and the mean performance is reported. The models are trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32. A learning rate reduction strategy is applied, where the learning rate is decreased by a factor of 0.1 every 30 epochs if the validation performance does not improve, promoting stable convergence. Early stopping is also employed to prevent overfitting based on validation loss. All other hyperparameters are determined empirically through cross-validation.

To evaluate the semi-supervised learning framework, we experiment with multiple unlabeled ratios. The unlabeled ratio denotes the proportion of training samples without ground-truth labels. In our experiments, we consider unlabeled ratios of 10%, 30%, 50%, 70%, and 90%, where only the remaining samples are treated as labeled data. We fix Top-$K = 5$ for hypergraph neighbor selection across both modalities, and tune the pseudo-label confidence threshold τ and the loss weight λ_u based on the unlabeled ratio and validation performance. To reduce implementation overhead, dynamic hyperedge construction is vectorized on GPU using L_2-normalized embeddings, batched similarity computation, and a GPU Top-K operation, avoiding explicit Python loops.

4.3. Evaluation metrics

To comprehensively evaluate the performance of the proposed HyperDyG architecture, we employ four standard metrics: Accuracy (Acc), Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1-score (WF1). These metrics jointly capture overall classification performance and class-level balance, which are particularly important in emotion recognition due to dataset imbalance.

WA is a weighted average of class-wise accuracies, with weights that increase the contribution of classes with larger sample sizes to the final score. In contrast, the UA measures the average accuracy across all emotion classes equally, mitigating the bias toward majority classes [35]. The formulations of WA and UA are expressed in Eqs. (21) and (22).

$$\mathrm{WA} = \sum_{i=1}^{K} w_i \cdot \frac{TP_i}{TP_i + FN_i}, \tag{21}$$

$$\mathrm{UA} = \frac{1}{K} \sum_{i=1}^{K} \frac{TP_i}{TP_i + FN_i}, \tag{22}$$

where TP, TN, FP, and FN denote the True Positives, True Negatives, False Positives, and False Negatives. Furthermore, K represents the total number of emotion classes, and w_i represents the weight assigned to class i.

The WF1 introduces class-specific weights w_i that reflect the class distribution to account for class imbalance. The formulation of WF1 is given in Eq. (23).

$$\mathrm{WF1} = \sum_{i=1}^{K} w_i \cdot \mathrm{F1}_i, \tag{23}$$

In addition to performance, we also report computational complexity in terms of the number of parameters, floating-point operations per second (FLOPs), and inference time. These measures collectively reflect the model's memory footprint, computational efficiency, and real-time feasibility during deployment.

5. Performance results

5.1. Experimental results of the HyperDyG architecture

The experimental result of the proposed HyperDyG confirms the effectiveness and robustness across multiple datasets and supervision settings. As shown in Tables 2 and 3, HyperDyG consistently outperforms the supervised baseline when pseudo-labeling is enabled, particularly at moderate unlabeled ratios. With supervised learning, the HyperDyG improves robustness with 82.53% on WA, 83.54% UA and 82.52% on WF1. The architecture also outperforms baselines on the ESD dataset, achieving 95.64% across the WA, UA, and WF1 metrics. A similarly high level of performance is observed on the ESD dataset, where HyperDyG reaches 95.64% across WA, UA, and WF1. These results indicate that the core architecture is competitive even without semi-supervised enhancements.

For the semi-supervised learning evaluation, the HyperDyG architecture shows further improvements, especially under mid-range unlabeled conditions. At 10% unlabeled ratio on the IEMOCAP dataset, WA increases by 2.46%, which goes from 79.42% to 81.88%, while UA improves by 2.03%. At larger unlabeled ratios, such as 70% and 90%, HyperDyG continues

Table 2. Performance comparison between supervised and semi-supervised learning under different unlabeled data ratios on the IEMOCAP dataset. **Bold values** denote the best overall performance across the entire table, while underlined values denote the best performance within each unlabeled-ratio pair. The (↑) indicates that higher values are better

Unlabeled ratio	Pseudo-labeling	Angry		Happy		Sad		Neutral		Overall performance		
		Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	WA (%) ↑	UA (%) ↑	WF1 (%) ↑
0%	✗	86.75	88.26	84.42	82.77	85.38	85.79	77.58	76.95	**82.53**	**83.54**	**82.52**
10%	✗	84.64	85.53	81.33	79.03	87.92	83.33	68.79	72.49	79.42	80.67	79.28
	✓	85.60	86.98	83.47	82.17	86.42	84.65	75.14	76.29	81.88	82.70	81.83
30%	✗	86.56	85.00	78.80	78.95	87.17	82.76	69.48	72.67	79.24	80.50	79.09
	✓	85.12	85.75	82.53	80.55	85.28	82.49	71.56	74.26	80.43	81.46	80.30
50%	✗	84.48	83.34	74.53	76.14	85.66	80.78	69.83	71.89	77.44	78.63	77.32
	✓	85.44	85.44	78.53	78.79	84.72	82.95	73.53	74.27	79.71	80.55	79.67
70%	✗	81.28	80.32	69.07	71.58	85.47	75.75	63.93	67.78	73.36	74.94	73.16
	✓	83.20	83.40	75.07	75.42	82.45	79.66	70.64	71.73	76.93	77.84	76.88
90%	✗	83.04	74.40	52.27	58.97	85.85	72.17	56.88	62.96	67.08	69.51	66.22
	✓	83.50	78.59	62.96	62.81	81.23	76.01	62.92	61.84	72.16	72.87	72.10

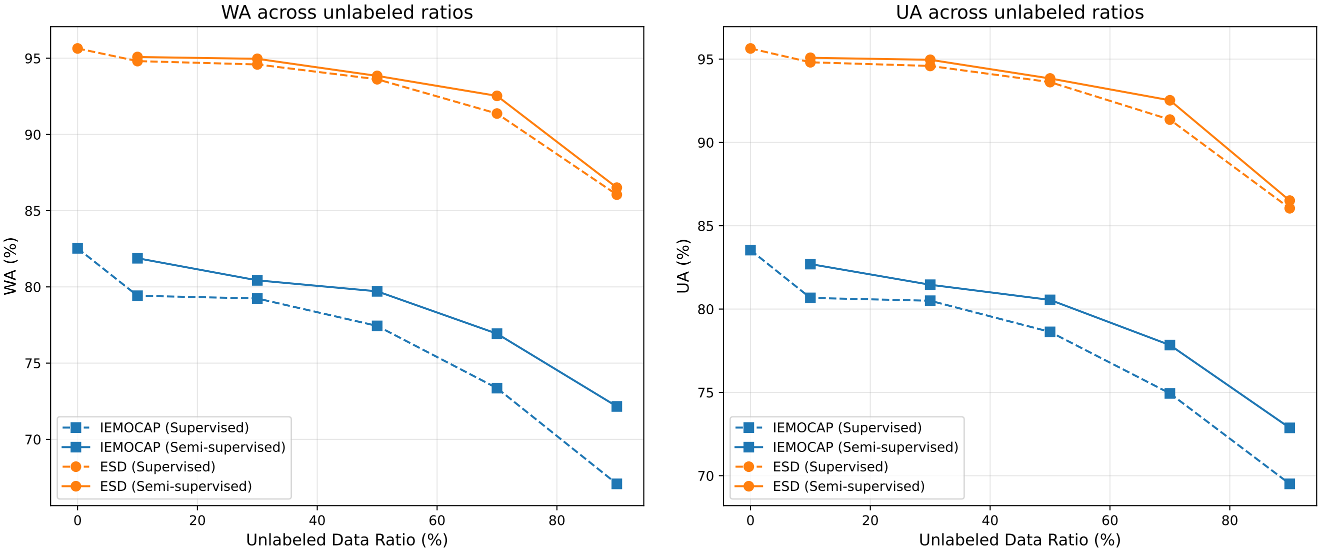
Table 3. Performance comparison between supervised and semi-supervised learning under different unlabeled data ratios on the ESD dataset. **Bold values** denote the best overall performance across the entire table, while underlined values denote the best performance within each unlabeled-ratio pair. The (↑) symbol denotes ahigher-is-better metric

Unlabeled ratio	Pseudo-labeling	Angry		Happy		Sad		Neutral		Suprise		Overall performance		
		Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	WA (%) ↑	UA (%) ↑	WF1 (%) ↑
0%	✗	96.94	95.89	93.68	93.20	97.15	97.44	98.10	97.44	93.76	94.66	**95.64**	**95.64**	**95.64**
10%	✗	96.34	95.36	91.37	91.76	97.14	96.76	96.74	97.05	92.46	93.10	94.81	94.81	94.81
	✓	95.54	94.95	93.03	92.21	97.23	96.74	97.43	97.40	92.17	93.45	95.08	95.08	95.07
30%	✗	95.37	94.86	91.89	91.60	96.91	96.50	96.57	97.01	92.23	93.00	94.59	94.59	94.59
	✓	96.03	94.89	92.25	92.86	96.43	96.80	97.74	97.26	92.35	93.70	94.96	94.96	94.96
50%	✗	94.91	93.95	89.20	90.10	95.66	95.63	96.17	96.36	92.17	92.04	93.62	93.62	93.62
	✓	96.01	94.25	92.31	90.89	95.88	95.92	96.67	96.38	91.30	92.82	93.84	93.84	93.85
70%	✗	93.71	92.35	84.74	86.62	95.89	93.93	93.94	95.19	88.57	88.70	91.37	91.37	91.35
	✓	96.14	93.23	87.28	88.30	95.38	95.29	96.60	96.16	91.75	90.35	92.53	92.53	92.52
90%	✗	92.11	89.08	73.43	78.06	90.46	89.56	91.03	91.00	83.20	81.89	86.05	86.05	85.94
	✓	94.48	89.66	78.68	79.73	92.80	79.73	91.77	91.10	82.06	82.81	86.51	86.51	86.42

to demonstrate strong resilience. Specifically, at 70% unlabeled data, WA rises by 3.57%, and at 90%, the improvement reaches 5.08%, indicating that the semi-supervised framework maintains stable gains even under extremely limited labeled conditions. Similarly, on the ESD dataset, HyperDyG exhibits consistent improvements across all unlabeled ratios. At lower unlabeled ratios, such as 10% and 30%, the performance improves by approximately 0.2–0.3% across WA, UA, and MF1 when comparing semi-supervised learning to the supervised baseline. At higher unlabeled ratios, the distance between the two methods becomes more significant. For example, at 70% unlabeled data, the gap reaches 1.16%, favoring the semi-supervised approach. This trend demonstrates that HyperDyG can effectively exploit unlabeled data and maintain stable learning dynamics even when the availability of labeled samples are severely limited.

Figure 3 also shows the trend of performance on IEMOCAP and ESD across the unlabeled ratio increases. On IEMOCAP, the supervised curves exhibit a clear downward slope as the unlabeled ratio increases. This shows the model is sensitive to the limited labeled data. In contrast, the semi-supervised curves decline at a significantly slower rate, indicating that pseudo-labeling and consistency regularization help stabilize learning. On the ESD dataset, both supervised and semi-supervised curves maintain relatively high accuracy due to the dataset's larger volume and cleaner distribution. However, the semi-supervised version consistently outperforms the supervised baseline across all ratios, with the gap becoming more visible at extreme ratios.

We also evaluate the model complexity of the HyperDyG architecture, which is presented in Table 4. The high computational cost originates from the pre-trained encoders, which contain WavLM and BERT. The parameters of the audio and text encoders reach 95.67M and 110.52M, respectively. In contrast, the learning model remains only 5.38M parameters and 0.08 GFLOPs, contributing minimally to the overall computational load. During inference, the learning model achieves the lowest latency of 3.15 ms/sample, while the total end-to-end processing time remains acceptable at 17.71 ms/sample. These results demonstrate that although



(a) WA across unlabeled ratios.

(b) UA across unlabeled ratios.

Figure 3. Performance comparison between supervised (dashed lines) and semi-supervised (solid lines) learning across different unlabeled data ratios for IEMOCAP and ESD datasets

Table 4. Model complexity analysis of the HyperDyG architecture and the encoder components. The (↓) indicates that lower values are better

Modules	**Model complexity**		
	Params ↓	GFLOPs ↓	Inference time ↓ (ms/sample)
WavLM	95.67M	0.34	7.41
BERT	110.52M	6.89	7.15
HyperDyG	5.38M	0.08	3.15
Total	211.57M	7.31	17.71

the full pipeline is dominated by high complexity in encoders, the proposed HyperDyG module introduces minimal overhead and is efficient enough for practical deployment scenarios.

HyperDyG performs well in both supervised and semi-supervised learning. The framework delivers solid baseline accuracy when trained solely on labeled data. The semi-supervised variant exhibits slower degradation under extreme unlabeled ratios, reduced sensitivity to label scarcity, and improved cross-modal consistency. The HyperDyG performance also maintains stronger cross-modal consistency. These behaviors confirm the robustness, scalability, and practical applicability of HyperDyG in real-world low-label scenarios. The decline in performance in the supervised setting is not severe across unlabeled ratios. This indicates that the core architecture remains relatively stable even as the proportion of labeled data decreases. However, the trade-off is that the performance gradually declines when labeled data becomes extremely scarce, although the semi-supervised strategy consistently mitigates this drop. In the future, we need to further enhance semi-supervised performance, particularly for a larger ratio of unlabeled data or more challenging data distributions.

5.2. Comparison between SOTA methods and the proposed approach

Table 5 presents a comprehensive comparison between the proposed HyperDyG framework and recent SOTA in MER approaches under supervised learning. On the IEMOCAP dataset, HyperDyG achieves 82.53% WA and 83.54% UA, outperforming strong competition approaches, such as Khan *et al.* [8], Fan *et al.* [41], and Wang *et al.* [42]. Compared with approaches that rely exclusively on cross-modal attention, such as MemoCMT, HyperDyG demonstrates clear advantages by improving WA by 0.68% and UA by 2.21%, confirming the benefits of integrating DHL with CMT for modeling higher-order multimodal dependencies. A similar trend is observed on the ESD dataset, where HyperDyG achieves 95.64% on both WA and UA, surpassing previously reported results and demonstrating strong generalization across datasets with different emotional distributions and recording conditions. Overall, these results confirm the robustness and superiority of the proposed HyperDyG framework over current SOTA approaches.

Table 5. Performance comparison of SER methods. **The bold font** denotes the best result, while the underlined font denotes the second-best result. In the modalities column, A represents audio and T represents text. The (*) denotes that this study employed only 4 emotions: anger, happiness, neutral, and sadness

References	Years	Modalities	IEMOCAP		ESD	
			WA (%) ↑	UA (%) ↑	WA (%) ↑	UA (%) ↑
Khan *et al.* [36]	2023	A	72.75	-	-	-
Prisayad *et al.* [37]	2023	A + T	76.80	77.30	-	-
Pham *et al.* [38]	2023	A + T	63.10	63.00	90.47*	90.46*
Wang *et al.* [26]	2023	A + T	75.20	76.40	-	-
Khurana *et al.*[39]	2024	A + T	73.00	73.00	-	-
Khan *et al.*[9]	2024	A + T	76.80	77.30	-	-
Kyung *et al.* [25]	2024	A + T	76.11	77.16	-	-
Yang *et al.* [40]	2024	A + T	-	-	88.50	88.50
Fan *et al.* [41]	2025	A + T	80.29	81.04	-	-
Qi *et al.* [14]	2025	A + T	77.30	78.20	-	-
Wang *et al.* [42]	2025	A + T	78.87	80.24	-	-
Khan *et al.* [8]	2025	A + T	81.85	81.33	91.84	91.93
HyperDyG	2025	A + T	**82.53**	**83.54**	**95.64**	**95.64**

Table 6. Evaluation of module contributions and configuration order effects on the performance of the HyperDyG model on the IEMOCAP dataset

Evaluations	Settings	Performance results		
		WA (%) ↑	UA (%) ↑	WF1 (%) ↑
HyperDyG	Original	**82.53**	**83.54**	**82.52**
Modules	w/o DHL	80.40	81.52	80.32
	w/o CMT	81.05	82.27	80.92
Graph learning	w/o R-GCN	81.62	82.56	81.58
	w/o Graph transformers	81.34	82.35	81.28
Configuration	DHL → CMT	80.32	81.25	80.32
	CMT → DHL	80.04	81.03	79.98

5.3. Ablation study

Impact of module and modality contributions. Table 6 presents the ablation analysis evaluating the contributions of individual modules and the influence of sequential configurations within the HyperDyG architecture. Removing either the DHL or the CMT modules leads to a noticeable performance degradation, confirming their complementary roles in capturing relational and contextual dependencies. Specifically, removing the DHL module reduces WA by 2.13% and WF1 by 2.20%, a greater decline than that observed when removing CMT, with a 1.48% and 1.60% drop in WA and WF1, respectively. This suggests that structural modeling via DHL is particularly critical for enhancing multimodal interactions and relational reasoning.

The graph learning components also contribute positively to feature refinement. Excluding the R-GCN decreases WA by 0.91%, while removing the Graph Transformer results in a 1.19% reduction, indicating that both modules jointly improve local and global dependency modeling. Moreover, when comparing sequential configurations, from the DHL to the CMT (DHL→CMT) and from the CMT to the DHL (CMT→DHL) variants show a substantial decrease of 2.21% and 2.49% in WA, respectively, compared to the parallel fusion design. This demonstrates that HyperDyG's dual-branch configuration with gated multimodal fusion achieves a more balanced integration of contextual and structural cues.

On the other hand, we evaluate the impact of modality on the HyperDyG architecture across the IEMOCAP and ESD datasets. In Table 7, the IEMOCAP integrating text with audio yields notable gains of 6.54% in WA and 6.60% in WF1 over the audio-only setting, reflecting the value of semantic context in conversational emotion disambiguation. In contrast,

Table 7. Performance comparison of the HyperDyG model under different modality configurations across two benchmark datasets

Datasets	Modality	Performance results		
		WA (%) ↑	UA (%) ↑	WF1 (%) ↑
IEMOCAP	Audio + Text	**82.53**	**83.54**	**82.52**
	Audio only	75.99	77.23	75.92
	Text only	55.42	53.73	53.74
ESD	Audio + Text	**95.64**	**95.64**	**95.64**
	Audio only	95.26	95.26	95.26
	Text only	20.00	20.00	6.67

on ESD, the improvement from multimodal fusion is marginal, suggesting that audio alone captures sufficient prosodic cues in short, acted utterances. Across all datasets, the text modality performs substantially worse, underscoring the limited expressiveness without prosodic support. These results highlight that while audio serves as the dominant modality for emotional salience, incorporating text enhances robustness and contextual understanding, particularly in spontaneous or linguistically rich scenarios.

Impact of supervised pre-training in semi-supervised framework. In Tables 8 and 9, we evaluate the role of supervised pre-training before the semi-supervised phase across different unlabeled data ratios. For the IEMOCAP dataset, the inclusion of supervised pre-training leads to performance improvements of approximately 0.5–1.5% across both metrics, particularly at mid and high unlabeled ratios. On the ESD dataset, supervised pre-training yields only modest gains, with the performance gap between the pre-trained and non-pre-trained models generally remaining within 0.4-0.9% for unlabeled ratios of 30–90%. At a 10% unlabeled ratio, the difference is negligible and marginally favors the non-pretrained model. Overall, incorporating supervised pre-training into the semi-supervised learning framework establishes a robust multimodal embedding space that effectively aligns the DHL and CMT branches. The approach mitigates noise amplification from low-confidence pseudo-labels, enabling the model to maintain stable performance under extreme unlabeled conditions.

Table 8. Impact of supervised pre-training on the IEMOCAP dataset for the semi-supervised HyperDyG framework across different unlabeled ratios. The underlined values denote the best performance within each unlabeled-ratio pair

Unlabeled ratio	Supervised pre-training	Overall performance		
		WA (%) ↑	UA (%) ↑	WF1 (%) ↑
10%	✗	81.37	82.40	81.27
	✓	81.88	82.70	81.83
30%	✗	80.03	80.83	80.17
	✓	80.43	81.46	80.30
50%	✗	78.30	79.05	78.27
	✓	79.71	80.55	79.67
70%	✗	76.04	76.85	76.03
	✓	76.93	77.84	76.88
90%	✗	71.08	72.01	70.93
	✓	72.16	72.87	72.10

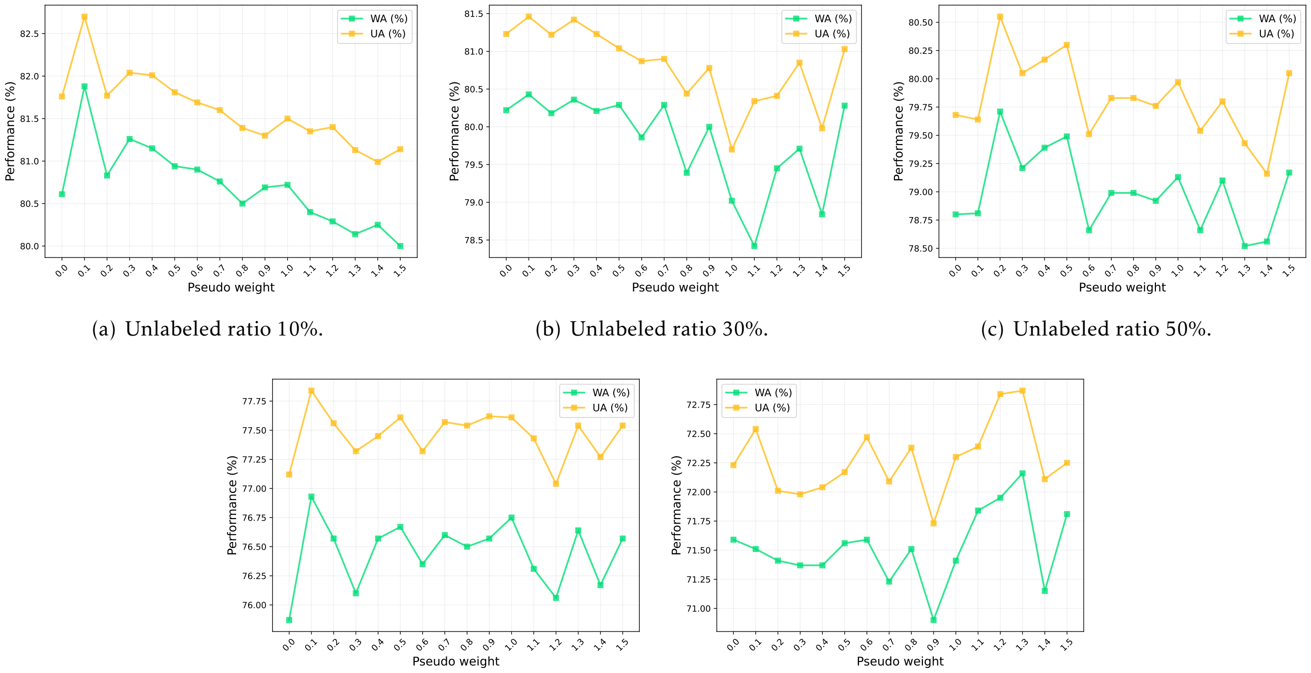
Table 9. Impact of supervised pre-training on the ESD dataset for the semi-supervised HyperDyG framework across different unlabeled ratios. The underlined values denote the best performance within each unlabeled-ratio pair

Unlabeled ratio	Supervised pre-training	Overall performance		
		WA (%) ↑	UA (%) ↑	WF1 (%) ↑
10%	✗	95.15	95.15	95.15
	✓	95.08	95.08	95.07
30%	✗	94.47	94.47	94.47
	✓	94.96	94.96	94.96
50%	✗	93.45	93.45	93.45
	✓	93.84	93.84	93.85
70%	✗	91.93	91.93	91.92
	✓	92.53	92.53	92.52
90%	✗	85.56	85.56	85.48
	✓	86.51	86.51	86.42

Impact of the balance coefficients in the loss function. The sensitivity in Figure 4 shows the interaction between the pseudo-label balance coefficient λ_u and the unlabeled ratio. At a low unlabeled ratio, the optimal coefficient is approximately 0.1, indicating that the $\mathcal{L}_{\text{semi}}$ must remain weak. Training on these cases follows the $\mathcal{L}_{\text{CE}}$ objective, as sufficient labeled data remain available. At a moderate ratio, the best performance is achieved when $\lambda_u = 0.3$, reflecting the increased contribution of pseudo-labeled samples and the need for stronger consistency regularization. At the extreme 90% unlabeled ratio, the optimal coefficient rises sharply to 1.3, demonstrating that the model must rely heavily on consistency constraints to compensate for the severely limited labeled data.

On the other hand, all five sensitivity plots consistently demonstrate the necessity of incorporating $\mathcal{L}_{\text{semi}}$. When $\lambda_u = 0$, meaning that $\mathcal{L}_{\text{semi}}$ is not applied, the performance remains at the lowest level across all settings. When the $\lambda_u > 0$, the accuracy curves increase noticeably, confirming that pseudo-label supervision and consistency regularization provide meaningful performance gains even when the unlabeled ratio varies.

In summary, the sensitivity plots consistently demonstrate the necessity of incorporating $\mathcal{L}_{\text{semi}}$. The analysis also highlights the crucial role of $\mathcal{L}_{\text{semi}}$ in stabilizing representations, suppressing noise in pseudo-labels, and maintaining coherent multimodal alignment when labeled data are scarce. These observations underscore the importance of striking a dynamic balance



(a) Unlabeled ratio 10%.

(b) Unlabeled ratio 30%.

(c) Unlabeled ratio 50%.

(d) Unlabeled ratio 70%.

(e) Unlabeled ratio 90%.

Figure 4. Sensitivity analysis of the balance coefficient associated with the IEMOCAP dataset with pseudo-label contributions in $\mathcal{L}_{total}$ under varying unlabeled data ratios

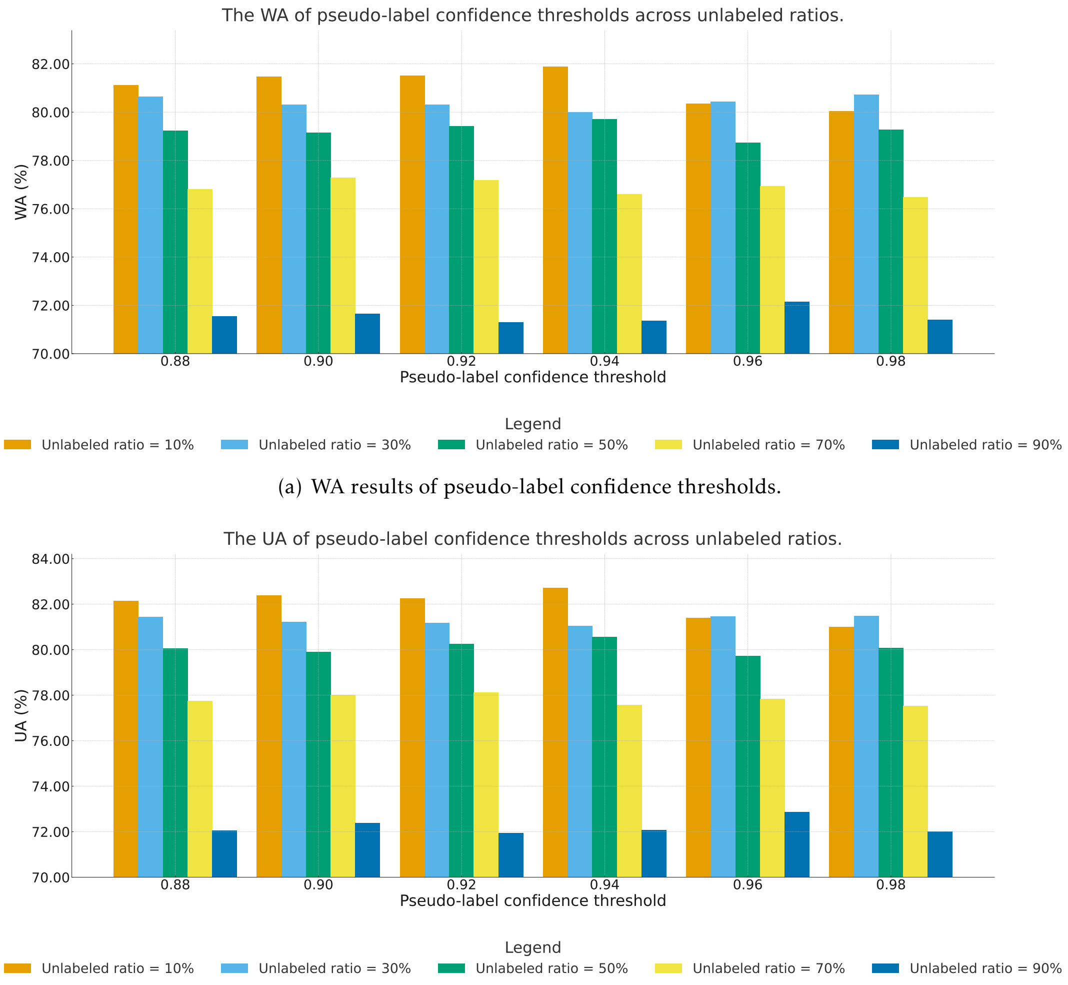
between supervised and semi-supervised objectives, adjusting the proportion of unlabeled data to fully leverage the benefits of hybrid learning. Nevertheless, at high unlabeled ratios, $\mathcal{L}_{\text{semi}}$ mitigates degradation but also amplifies pseudo-label noise if overweighted, emphasizing the need for carefully adjusted balance coefficients and pseudo-label thresholds.

Impact of pseudo-label confidence threshold. For the semi-supervised learning framework, the confidence threshold τ is a critical hyperparameter that governs the trade-off between the quality of pseudo-labels and their coverage. We sweep $\tau \in [0.88, 0.98]$ under unlabeled ratios with identical training settings and report performance in Figure 5. At lower unlabeled ratios, which range from 10% to 50%, performance peaks near $\tau \approx 0.94$ and declines for larger values. Under sufficient labeled supervision, excessive threshold stringency prunes valuable unlabeled samples, suppresses acceptance, and diminishes accuracy. At higher unlabeled ratios from 70% to 90%, the optimal threshold shifts upward to 0.96, reflecting the need to filter noisier pseudo-labels when labeled guidance is scarce. However, excessively large τ again collapses acceptance and harms accuracy.

The observed trend is consistent with known limitations of FixMatch-style self-training. As labeled supervision weakens, pseudo-labels become noisier, and incorrect high-confidence predictions are more easily reinforced, amplifying errors. In the setting, the τ acts as a minimal noise-control mechanism by suppressing unreliable pseudo-labels, while an overly conservative τ prevents the model from effectively leveraging unlabeled data. These findings motivate more robust adaptive strategies, including scheduling τ according to uncertainty and training dynamics. Moreover, integrating confidence calibration and teacher-based stabilization can further improve robustness under large unlabeled ratios.

Impact of fusion strategies in HyperDyG architecture. Figure 6 illustrates the comparative performance of various fusion strategies integrated into the HyperDyG framework on the IEMOCAP dataset. Among all approaches, the GMU achieves the highest accuracy and F1-scores, demonstrating its superior ability to dynamically regulate modality contributions based on contextual relevance. Traditional methods, such as simple concatenation or element-wise operations, exhibit inferior performance due to their inability to adaptively weight the importance of modalities.

Although static fusion schemes are simpler and more computationally efficient, the learnable gating mechanism in GMU provides fine-grained control over multimodal interactions, yielding more discriminative emotional representations. This adaptive strategy



(a) WA results of pseudo-label confidence thresholds.

(b) UA results of pseudo-label confidence thresholds.

Figure 5. Sensitivity to pseudo-label confidence threshold τ on IEMOCAP dataset

enables HyperDyG to effectively balance structural cues from DHL and contextual cues from CMT, forming a unified feature space that enhances robustness and generalization in MER.

Impact of supervised and semi-supervised HyperDyG framework under noisy conditions. Table 10 presents the performance of the HyperDyG architecture under various NoiseX-92 environments [43] at different Signal-to-Noise Ratios (SNRs). Across all noise types, the supervised model exhibits gradual degradation as noise intensity increases, particularly under low SNRs (0–10 dB), where acoustic distortion significantly affects emotional cues. However, the semi-supervised configurations often outperform the purely supervised baseline, especially at moderate SNR levels, demonstrating enhanced robustness through the utilization of unlabeled data.

Among the tested noises, the Babble and Volvo conditions cause the most substantial degradation due to overlapping speech and engine harmonics, which interfere with temporal dynamics. In contrast, HF Channel and White noise show relatively smaller performance gaps, indicating HyperDyG's resilience to stationary disturbances. Notably, even with 50%–70% unlabeled data ratios, the semi-supervised variant maintains stable WA and UA scores, confirming the model's capacity to learn noise-invariant features from weakly supervised signals. These findings highlight that integrating semi-supervised learning not only alleviates data scarcity but also improves generalization and robustness against real-world acoustic perturbations.

5.4. Limitation and discussion

In our experiments, HyperDyG achieves competitive supervised results and maintains relatively stable performance as the unlabeled ratio increases. The dual-branch architecture within DHL for higher-order structure and CMT for contextual alignment provides complementary cues. The GMU fusion improves

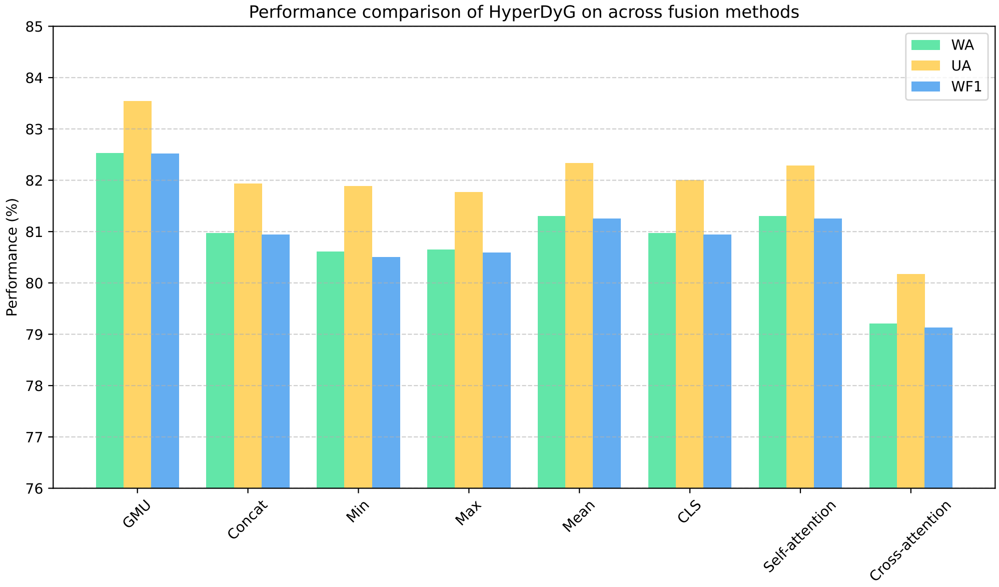


Figure 6. The performance of the HyperDyG model on the IEMOCAP dataset across different fusion methods

Table 10. Performance result of the HyperDyG architecture on the IEMOCAP dataset under each NoiseX-92 type

Noise type	SNRs (dB)	Supervised		SSL (10% U)		SSL (30% U)		SSL (50% U)		SSL (70% U)		SSL (90% U)	
		WA (%) ↑	UA (%) ↑	WA (%) ↑	UA (%) ↑	WA (%) ↑	UA (%) ↑	WA (%) ↑	UA (%) ↑	WA (%) ↑	UA (%) ↑	WA (%) ↑	UA (%) ↑
Babble	20	74.55	75.46	75.09	76.05	72.38	72.34	69.13	70.29	67.87	68.55	60.11	62.29
	15	69.68	69.68	68.95	69.97	68.77	68.59	63.54	64.31	62.82	62.99	56.32	57.99
	10	61.37	61.37	58.66	59.81	59.75	58.95	54.87	54.67	53.25	52.62	47.65	48.71
	0	38.09	38.09	39.89	41.64	41.16	41.02	38.99	36.71	39.89	37.87	35.02	35.89
F16	20	66.97	67.91	66.43	67.87	63.90	64.39	61.19	63.51	60.65	62.21	51.99	54.32
	15	60.38	61.55	58.12	59.40	57.40	57.77	54.15	56.29	51.08	52.85	40.25	43.06
	10	51.44	52.19	48.92	49.95	48.74	49.29	42.96	45.18	40.25	42.39	31.41	34.85
	0	31.41	33.69	30.51	33.63	30.32	33.18	26.90	30.56	24.01	24.01	21.84	27.17
HF Channel	20	65.16	64.38	64.98	64.32	63.90	62.47	63.54	64.29	64.80	65.94	57.94	59.82
	15	60.47	59.10	62.45	61.08	60.29	58.75	61.19	61.36	61.19	61.71	56.32	57.76
	10	57.22	55.35	57.04	55.01	55.42	53.05	58.12	57.11	57.76	56.97	54.15	54.66
	0	38.45	33.60	39.53	34.65	37.91	32.40	39.53	34.67	42.24	37.66	43.32	40.80
Volvo	20	75.09	75.99	74.91	75.99	73.83	73.66	71.30	72.54	68.41	69.30	63.90	66.11
	15	73.47	74.45	73.65	74.87	73.10	73.22	70.40	71.77	67.87	69.06	62.64	65.01
	10	71.84	73.02	71.12	72.76	72.38	72.68	67.69	69.59	66.43	68.13	57.40	60.37
	0	65.34	67.80	64.08	66.77	64.08	65.74	57.58	61.23	54.69	58.16	43.50	48.46
White	20	66.97	67.71	65.52	66.69	64.08	64.80	62.09	64.50	62.82	64.68	56.32	58.93
	15	63.36	63.40	62.82	63.34	59.93	60.10	60.11	62.25	59.75	61.67	55.23	57.47
	10	60.11	59.29	60.29	59.84	56.50	55.74	58.12	59.51	54.33	55.8	51.44	53.14
	0	40.43	37.64	40.97	38.46	38.81	36.70	38.45	37.53	37.91	38.05	34.66	35.93

consistent gains over static fusion baselines. We further observe robustness under acoustic noise perturbations in the evaluated settings, suggesting practical potential in adverse conditions.

Due to the limitations of the HyperDyG architecture, the pseudo-label balance coefficient λ_u in the total loss is fixed rather than dynamically adapted to the training dynamics, data difficulty, or calibration status. As a result, the static weighting can become misaligned across different label ratios and epochs, leading to suboptimal performance. The modality embedding extractors dominate parameters and FLOPs, thereby inflating the training and inference costs. Our current evaluation focuses on the audio-text setting and does not include the visual modality. In future work, we will extend HyperDyG to tri-modal fusion and validate it with visual signals. Finally, like most self-training pipelines, error propagation from noisy pseudo-labels remains a risk, particularly for minority classes and at extreme unlabeled ratios. Moreover, mean pooling in hyperedge aggregation is lightweight but may under-model node importance and subset interactions. We will explore attention-based pooling in future work.

In future work, we plan to make the pseudo-label balance λ_u and the confidence threshold τ adaptive, driven by uncertainty- and curriculum-based signals [44–46]. The technology can be applied, such as temperature-scaled calibration, agreement-based filtering [47, 48], and the exponential moving average teacher ramps [49, 50]. We will further explore dynamic neighborhood selection in DHL. To reduce encoder complexity, we will investigate knowledge distillation [51], structured pruning with post-training quantization [52], and lighter backbones, aiming to preserve accuracy under tighter compute budgets. We will incorporate reliability-aware consistency losses, prototype refinement, and co- and tri-training to mitigate pseudo-label drift. These directions aim to retain HyperDyG's benefits while improving efficiency, robustness, and scalability.

6. Conclusion

In this work, we introduce HyperDyG, a dynamic hypergraph-driven MER framework that integrates semi-supervised learning to effectively leverage both labeled and unlabeled data. The proposed dual-stream architecture combines the DHL and CMT modules together with a novel GMU fusion mechanism. HyperDyG achieves successful structural dependencies, contextual alignment, and content-specific interactions across audio and text modalities. The dual-stream modeling enables the architecture to construct richer relational representations while preserving fine-grained cross-modal coherence. The HyperDyG achieves SOTA on IEMOCAP and ESD, outperforming strong multimodal baselines in both supervised and semi-supervised settings and demonstrating robustness under acoustic noise. These findings substantiate the efficacy of combining dynamic hypergraph reasoning with attention-based alignment and gated fusion for scalable MER. In the future, we need to enhance HyperDyG with adaptive pseudo-labeling, dynamic neighborhood selection in DHL, and efficiency-oriented encoders to improve scalability without sacrificing accuracy.

Abbreviations

Abbreviation	Definition
Acc	Accuracy
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
CH-GAT	Cross-modal Heterogeneous Graph Attention Network
CMT	Cross-Modal Transformer
DHL	Dynamic Hypergraph Learning
ESD	Emotional Speech Dataset
F1	F1-score
GFLOPs	Giga Floating-Point Operations Per Second
GMU	Gated Multimodal Unit
GPU	Graphic Processing Unit
HCI	Human-Computer Interaction
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LLM	Large Language Model
MER	Multimodal Emotion Recognition
R-GCN	Relational Graph Convolutional Network
SER	Speech Emotion Recognition
SNRs	Signal-to-Noise Ratios
SOTA	State-of-the-art
UA	Unweighted Accuracy
WA	Weighted Accuracy
WF1	Weighted F1-score
xCBAM	Cross-modal Convolutional Block Attention Mechanism

Data availability

In this manuscript, we use publicly available datasets, including IEMOCAP and ESD, obtained from their respective sources. We provide additional details on dataset access and preprocessing procedures in the manuscript. Our source code is publicly available at https://github.com/nhut-ngnn/HyperDyG.

References

[1] Liu ZT, Rehman A, Wu M, Cao WH, Hao M. Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. Information Sciences. 2021;563:309-25. Available from: https://doi.org/10.1016/j.ins.2021.02.016.

[2] Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access. 2021;9:47795-814. Available from: https://doi.org/10.1109/ACCESS.2021.3068045.

[3] George SM, Muhamed Ilyas P. A review on speech emotion recognition: A survey, recent advances,

challenges, and the influence of noise. Neurocomputing. 2024;568:127015. Available from: https://doi.org/10.1016/j.neucom.2023.127015.

[4] Pudasaini A, Al-Hawawreh M, Bouadjenek MR, Hacid H, Aryal S. A comprehensive study of audio profiling: Methods, applications, challenges, and future directions. Neurocomputing. 2025;640:130334. Available from: https://doi.org/10.1016/j.neucom.2025.130334.

[5] Ahmed N, Aghbari ZA, Girija S. A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications. 2023;17:200171. Available from: https://doi.org/10.1016/j.iswa.2022.200171.

[6] Hazmoune S, Bougamouza F. Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. Engineering Applications of Artificial Intelligence. 2024;133:108339. Available from: https://doi.org/10.1016/j.engappai.2024.108339.

[7] Nguyen NM, Nguyen TT, Tran PN, Lim CP, Pham NT, Dang DNM. Multimodal fusion in speech emotion recognition: A comprehensive review of methods and technologies. Engineering Applications of Artificial Intelligence. 2026;163:112624. Available from: https://doi.org/10.1016/j.engappai.2025.112624.

[8] Khan M, Tran PN, Pham NT, El Saddik A, Othmani A. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. Scientific reports. 2025;15(1):5473. Available from: https://doi.org/10.1038/s41598-025-89202-x.

[9] Khan M, Gueaieb W, El Saddik A, Kwon S. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. Expert Systems with Applications. 2024;245:122946. Available from: https://doi.org/10.1016/j.eswa.2023.122946.

[10] Xie Y, Sun C, Cao Z, Liu B, Ji Z, Liu Y, et al. A Dual Contrastive Learning Framework for Enhanced Multimodal Conversational Emotion Recognition. In: Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics; 2025. p. 4055-65.

[11] Xiang J, Zhu X, Cambria E. Integrating audio–visual text generation with contrastive learning for enhanced multimodal emotion analysis. Information Fusion. 2026;127:103809. Available from: https://doi.org/10.1016/j.inffus.2025.103809.

[12] Nguyen LH, Pham NT, Khan M, Othmani A, EI Saddik 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. MMAsia ’24. New York, NY, USA: Association for Computing Machinery; 2024. p. 1 6. Available from: https://doi.org/10.1145/3696409.3700183.

[13] Nguyen NM, Le TT, Nguyen TT, Phan DT, Tran AK, Dang DNM. CemoBAM: Advancing Multimodal Emotion Recognition through Heterogeneous Graph Networks and Cross-Modal Attention Mechanisms. In: 2025 25th Asia-Pacific Network Operations and Management Symposium (APNOMS); 2025. p. 1-4. Available from: https://doi.org/10.23919/APNOMS67058.2025.11181320.

[14] Qi X, Wen Y, Zhang P, Huang H. MFGCN: Multimodal fusion graph convolutional network for speech emotion recognition. Neurocomputing. 2025;611:128646. Available from: https://doi.org/10.1016/j.neucom.2024.128646.

[15] Fan C, Lin J, Mao R, Cambria E. Fusing pairwise modalities for emotion recognition in conversations. Information Fusion. 2024;106:102306. Available from: https://doi.org/10.1016/j.inffus.2024.102306.

[16] Zhang S, Chen M, Chen J, Li YF, Wu Y, Li M, et al. Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition. Knowledge-Based Systems. 2021;229:107340. Available from: https://doi.org/10.1016/j.knosys.2021.107340.

[17] Hady MFA, Schwenker F. In: Semi-supervised Learning. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 215-39. Available from: https://doi.org/10.1007/978-3-642-36657-4_7.

[18] Yang X, Song Z, King I, Xu Z. A Survey on Deep Semi-Supervised Learning. IEEE Transactions on Knowledge and Data Engineering. 2023;35(9):8934-54. Available from: https://doi.org/10.1109/TKDE.2022.3220219.

[19] Arazo E, Ortego D, Albert P, O’Connor NE, McGuinness K. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In: 2020 International Joint Conference on Neural Networks (IJCNN); 2020. p. 1-8. Available from: https://doi.org/10.1109/IJCNN48605.2020.9207304.

[20] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. p. 1195 1204.

[21] Xie Q, Luong MT, Hovy E, Le QV. Self-Training With Noisy Student Improves ImageNet Classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 10684-95. Available from: https://doi.org/10.1109/CVPR42600.2020.01070.

[22] Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, et al. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In: Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc.; 2020. p. 596-608.

[23] Agarla M, Bianco S, Celona L, Napoletano P, Petrovsky A, Piccoli F, et al. Semi-supervised cross-lingual speech emotion recognition. Expert Systems with Applications. 2024;237:121368. Available from: https://doi.org/10.1016/j.eswa.2023.121368.

[24] Chen H, Guo C, Li Y, Zhang P, Jiang D. Semi-Supervised Multimodal Emotion Recognition with Class-Balanced Pseudo-labeling. In: Proceedings of the 31st ACM International Conference on Multimedia. MM ’23. New York, NY, USA: Association for Computing Machinery; 2023. p. 9556–9560. Available from: https://doi.org/10.1145/3581783.3612864.

[25] Kyung J, Heo S, Chang JH. Enhancing Multimodal Emotion Recognition through ASR Error Compensation and LLM Fine-Tuning. In: Interspeech 2024; 2024. p. 4683-7. Available from: https://doi.org/10.

21437/Interspeech.2024-2364.

[26] Wang S, Ma Y, Ding Y. Exploring Complementary Features in Multi-Modal Speech Emotion Recognition. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023. p. 1-5. Available from: https://doi.org/10.1109/ICASSP49357.2023.10096709.

[27] Tsouvalas V, Ozcelebi T, Meratnia N. Privacy-preserving Speech Emotion Recognition through Semi-Supervised Federated Learning. In: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); 2022. p. 359-64. Available from: https://doi.org/10.1109/PerComWorkshops53856.2022.9767445.

[28] Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing. 2022;16(6):1505-18. Available from: https://doi.org/10.1109/JSTSP.2022.3188113.

[29] Koroteev MV. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:210311943. 2021.

[30] Feng Y, You H, Zhang Z, Ji R, Gao Y. Hypergraph Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul;33(01):3558-65. Available from: https://doi.org/10.1609/aaai.v33i01.33013558.

[31] Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling Relational Data with Graph Convolutional Networks. In: The Semantic Web. Cham: Springer International Publishing; 2018. p. 593-607. Available from: https://doi.org/10.1007/978-3-319-93417-4_38.

[32] Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph Transformer Networks. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc.; 2019. p. 11983 11993. Available from: https://doi.org/10.5555/3454287.3455360.

[33] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008;42:335-59. Available from: https://doi.org/10.1007/s10579-008-9076-6.

[34] Zhou K, Sisman B, Liu R, Li H. Emotional voice conversion: Theory, databases and ESD. Speech Communication. 2022;137:1-18. Available from: https://doi.org/10.1016/j.specom.2021.11.006.

[35] Liu S, Gao P, Li Y, Fu W, Ding W. Multi-modal fusion network with complementarity and importance for emotion recognition. Information Sciences. 2023;619:679-94. Available from: https://doi.org/10.1016/j.ins.2022.11.076.

[36] Khan M, El Saddik A, Alotaibi FS, Pham NT. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. Knowledge-Based Systems. 2023;270:110525. Available from: https://doi.org/10.1016/j.knosys.2023.110525.

[37] Prisayad D, Fernando T, Sridharan S, Denman S, Fookes C. Dual Memory Fusion for Multimodal Speech Emotion Recognition. In: Interspeech 2023; 2023. p. 4543-7. Available from: https://doi.org/10.21437/Interspeech.2023-1090.

[38] Pham NT, Phan LT, Dang DNM, Manavalan B. 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. SOICT '23. New York, NY, USA: Association for Computing Machinery; 2023. p. 870–877. Available from: https://doi.org/10.1145/3628797.3628887.

[39] Khurana Y, Gupta S, Sathyaraj R, Raja SP. RobinNet: A Multimodal Speech Emotion Recognition System With Speaker Recognition for Social Interactions. IEEE Transactions on Computational Social Systems. 2024;11(1):478-87. Available from: https://doi.org/10.1109/TCSS.2022.3228649.

[40] Yang J, Liu J, Huang K, Xia J, Zhu Z, Zhang H. Single- and Cross-Lingual Speech Emotion Recognition Based on WavLM Domain Emotion Embedding. Electronics. 2024;13(7):1380. Available from: https://doi.org/10.3390/electronics13071380.

[41] Fan W, Xu X, Liu F, Xing X. Multimodal speech emotion recognition via dynamic multilevel contrastive loss under local enhancement network. Expert Systems with Applications. 2025;281:127669. Available from: https://doi.org/10.1016/j.eswa.2025.127669.

[42] Wang X, Zhao S, Sun H, Wang H, Zhou J, Qin Y. Enhancing Multimodal Emotion Recognition through Multi-Granularity Cross-Modal Alignment. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025. p. 1-5. Available from: https://doi.org/10.1109/ICASSP49660.2025.10889156.

[43] Varga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication. 1993;12(3):247-51. Available from: https://doi.org/10.1016/0167-6393(93)90095-3.

[44] Guo LZ, Li YF. Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding. In: Proceedings of the 39th International Conference on Machine Learning. vol. 162 of Proceedings of Machine Learning Research. PMLR; 2022. p. 8082-94. Available from: https://proceedings.mlr.press/v162/guo22e.html.

[45] Dong H, Rodríguez AM, Guinaudeau C, Satoh S. Fairness Without Labels: Pseudo-Balancing for Bias Mitigation in Face Gender Classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops; 2025. p. 7683-92.

[46] Zhang B, Wang Y, Hou W, WU H, Wang J, Okumura M, et al. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In: Advances in Neural Information Processing Systems. vol. 34. Curran Associates, Inc.; 2021. p. 18408-19.

[47] Manna S, Chattopadhyay S, Dey R, Pal U, Bhattacharya S. Dynamically Scaled Temperature in Self-Supervised Contrastive Learning. IEEE Transactions on Artificial Intelligence. 2025;6(6):1502-12. Available from: https://doi.org/10.1109/TAI.2024.3524979.

[48] Sanchez Aimar E, Helgesen N, Xu Y, Kuhlmann M, Felsberg M. Flexible Distribution Alignment: Towards Long-Tailed Semi-supervised Learning with Proper Calibration. In: Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 307-27. Available from: https://doi.org/10.1007/978-3-031-72949-2_18.

[49] Wang S, Sun X, Chen C, Hong D, Han J. Semi-Supervised Semantic Segmentation for Remote Sensing Images via Multiscale Uncertainty Consistency and Cross-Teacher–Student Attention. IEEE Transactions on Geoscience and Remote Sensing. 2025;63:1-15. Available from: https://doi.org/10.1109/TGRS.2025.3585489.

[50] Lilja A, Wallin E, Fu J, Hammarstrand L. Exploring Semi-Supervised Learning for Online Mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2025. p. 2502-12. Available from: https://www.doi.org/10.1109/CVPRW67362.2025.00233.

[51] Li S, Zhang T, Chen CLP. Cyclic Data Distillation Semi-Supervised Learning for Multi-Modal Emotion Recognition. IEEE Transactions on Knowledge and Data Engineering. 2025;37(9):5078-92. Available from: https://doi.org/10.1109/TKDE.2025.3581786.

[52] Liu B, Gu T, Wang H, Qian Y. MixPQ: Joint Pruning and Quantization for Speech and Language Foundation Models Compression. IEEE Transactions on Audio, Speech and Language Processing. 2025;33:4098-112. Available from: https://doi.org/10.1109/TASLPRO.2025.3613948.