ProgAttn-DAG: Progressive Attention-Based Modality Fusion in Directed Acyclic Graphs for Multimodal Conversational Emotion Recognition

Yitong Yao, Xiao Sun†
School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
2024110509@mail.hfut.edu.cn, sunx@hfut.edu.cn

***Abstract*—Affective computing constitutes a crucial branch of Natural Language Processing (NLP). Multimodal Emotion Recognition in Conversation (MERC) has emerged as a prominent task in multimodal emotion computing. It aims to process and fuse information from heterogeneous modalities (visual, acoustic, textual) to accurately recognize speakers' emotional states in conversations. Despite extensive research in emotion recognition, challenges persist. First, multimodal methods often employ simplistic single-step alignment and fixed-weight fusion strategies, inadequately addressing heterogeneity and failing to capture comprehensive cross-modal interactions. Second, conversational emotion recognition requires modeling complex contextual dependencies, as emotions evolve dynamically throughout dialogues. Third, benchmark datasets exhibit severe class imbalance with long-tail distributions, where minority emotion categories are underrepresented. To address these challenges, this paper proposes a novel Progressive Attention-Based Modality Fusion in Directed Acyclic Graphs (ProgAttn-DAG) framework for multimodal conversational emotion recognition. The framework integrates a progressive attention mechanism (ProgAttn) that hierarchically fuses text, audio, and visual modalities to address heterogeneity, multi-layer directed acyclic graph networks to explicitly model contextual dependencies, and curriculum learning strategies to optimize sample-difficulty-based training and mitigate data imbalance. Extensive experiments on two benchmark datasets demonstrate the effectiveness of ProgAttn-DAG, particularly in addressing data imbalance issues.**

***Index Terms*—Multimodal Emotion Recognition, Multimodal Fusion, Graph Neural Networks, Directed Acyclic Graph, Curriculum Learning**

I. Introduction

In psychology, emotion is defined as a multidimensional manifestation of human cognitive mechanisms [1], involving dynamic interactions between physiological arousal and cognitive assessment [2]. These interactions shape subjective emotional experiences, trigger cognitive processing, and regulate the autonomic nervous system, supporting adaptive behavioral responses. Therefore, affective computing plays a crucial role in developing artificial intelligence systems capable of emotional understanding and commonsense reasoning.

With the rise of social media, Multimodal Emotion Recognition in Conversation (MERC) has become a key research focus. Its applications include social media opinion mining, mental health diagnostics, intelligent customer service, and enhanced human-computer interaction. Despite significant progress, Emotion Recognition in Conversation (ERC) still faces several critical challenges.

First, most existing approaches rely solely on textual modality, limiting the ability to capture the diversity and complexity of emotional expressions. As shown in Fig. 1, the utterance "See? It's not a big deal." may be misclassified as neutral using only text, while visual cues (frowning, sighing, and downward mouth movements) reveal the true emotional state.

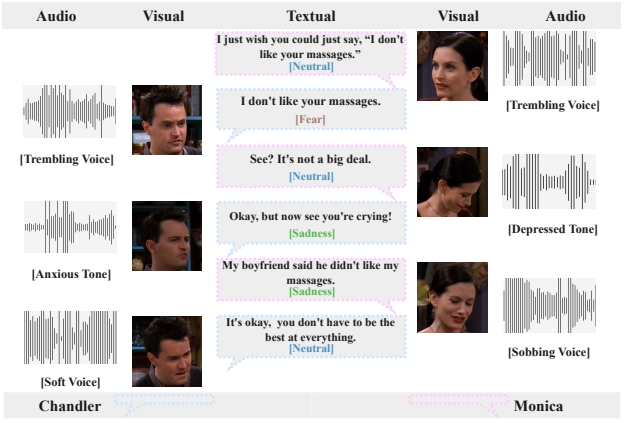

Fig. 1. Illustration of a multimodal conversation in the MELD dataset.

Second, even multimodal methods often employ simplistic single-step alignment and fixed-weight fusion strategies, which inadequately address modality heterogeneity. Additionally, data imbalance with long-tail effects persists in standard ERC benchmark datasets, where common emotion categories (e.g., anger, happiness, sadness) dominate, while minority categories (e.g., surprise, disgust) are severely underrepresented. To address these challenges, we propose ProgAttn-DAG, a novel multimodal fusion framework based on progressive attention mechanisms and directed acyclic graph (DAG) modeling for conversational emotion recognition.

To summarize, the main contributions of this paper can be summarized as follows:

- We propose ProgAttn-DAG, a novel framework that

This work was supported by Anhui Province Key R&D Program (ss202304a05020068), Special Project of the National Natural Science Foundation of China (62441614) and General Program of the National Natural Science Foundation of China (62376084).

†Corresponding author.

integrates text, audio, and visual modalities through a progressive attention mechanism, effectively addressing modality heterogeneity and capturing fine-grained cross-modal interactions.
- We employ Directed Acyclic Graph (DAG) modeling to represent dependency relationships among modalities and utterances, enabling comprehensive multimodal fusion and contextual modeling.
- We utilize curriculum learning strategies to optimize sample-difficulty-based training and mitigate data imbalance.
- Extensive experiments on two benchmark datasets demonstrate the effectiveness of ProgAttn-DAG, particularly in improving recognition of minority emotion categories and addressing data imbalance issues.

II. Related Work

In this section, we categorize existing approaches according to the input data modalities and systematically review unimodal emotion recognition methods, multimodal emotion recognition frameworks, and Curriculum Learning strategies.

A. Unimodal Emotion Recognition

Unimodal emotion recognition uses a single modality for affect analysis, with NLP mainly relying on text. Sequence-based methods include DialogueRNN [3]. To overcome sequence model limitations, graph methods emerged. DialogueGCN [4] models contextual dependencies; ConGCN [5] applies structure-constrained graph convolution; DAG-ERC [6] combines graph neural networks with recursive architectures.

B. Multimodal Emotion Recognition

Emotions conveyed in human conversations represent complex and elusive dynamic psychological phenomena. Therefore, the integration of multiple modalities plays a pivotal role in enhancing the accuracy of emotion recognition among participants in conversations. For modality fusion, COGMEN [7] and MMGCN [8] advanced fusion with graph neural networks and transformers. These studies show multimodal emotion recognition outperforms unimodal methods, yet most work focuses on architectural optimization while overlooking modality heterogeneity and data imbalance. Recent benchmarking further analyzes modality-level emotion conflicts and proposes protocols to evaluate and bridge them, complementing our fusion and curriculum strategies [9].

C. Curriculum Learning

Curriculum learning (CL) was introduced by Bengio et al. [10] as a training strategy that mimics human learning by progressively adjusting sample difficulty. Mao et al. [11] integrated emotional transition frequency difficulty meters and training schedulers to improve model discrimination.

III. Methodology

A. Problem Formulation

Given M dialogue sequences $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$, each dialogue c_i contains utterances $c_i = \{u_1, u_2, \ldots, u_n\}$, where n is the number of utterances. Each utterance (unique speaker in multi-party settings) occurs in a time interval and has visual (u_i^v), textual (u_i^t), and acoustic (u_i^a) features: $u_i = \{u_i^v, u_i^t, u_i^a\}$. The objective is to predict for each u_i an emotion label $e \in \mathcal{E} = \{e_1, \ldots, e_r\}$ without prior knowledge of speaker identities or number of participants.

B. Progressive Attention-Based Modality Fusion in Directed Acyclic Graphs

The architecture of our proposed ProgAttn-DAG model is illustrated in Fig. 2. In the following sections, we mainly present a detailed description of four key components that constitute our framework: Multimodal Feature Representation Learning, Multimodal Fusion, Multi-layer DAG Networks, and CL Strategies.

C. Multimodal Feature Representation Learning

In this section, we extract features for each modality through feature extraction, modality encoding, and contextual modeling. After optimizing individual modality representations, we perform multimodal fusion in Section 3.4 for comprehensive integration.

Visual Modality: For visual feature representation learning, we adopt ResNet101 [12] (among common pre-trained DenseNet [13] and 3D-CNN [14]) for its performance, training speed, and deployment efficiency. To mitigate background noise, we apply the multi-task MTCNN for efficient face region extraction, reducing irrelevant interference and improving recognition accuracy [15]. Using P-Net to propose candidate face boxes, R-Net to refine and eliminate overlaps, and O-Net to locate key landmarks for transformation parameters, we obtain the final aligned face region. Specifically, MTCNN detects faces in each video frame, and ResNet101 extracts 1000-dimensional features, which are refined by DialogueRNN to obtain 256-dimensional contextual visual representations, denoted as y_i^v.

Textual Modality: For textual feature representation learning, we adopt the One-stage paradigm [16] for efficiency. The framework models contextual information using three segments [17]: the historical context $\{u_1^t, u_2^t, \ldots, u_{i-1}^t\}$, the target utterance u_i^t, and the subsequent context $\{u_{i+1}^t, \ldots, u_n^t\}$. These segments are concatenated with [SEP] tokens and processed through a pre-trained RoBERTa [18] model and fully connected layers to generate 768-dimensional textual representations, which are further refined by DialogueRNN to obtain 256-dimensional contextual embeddings, denoted as y_i^t.

Audio Modality: For audio feature representation learning, we employ the open-source OpenSMILE toolkit [19] to extract 6,373-dimensional acoustic feature vectors from each utterance. These high-dimensional features are compressed to 512 dimensions via fully connected layers, and further refined

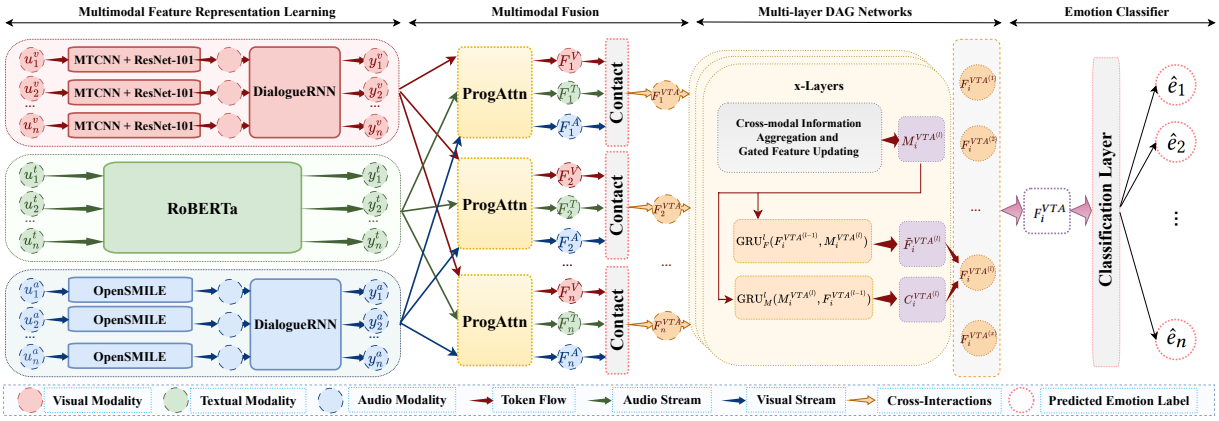

Fig. 2. An overview of our proposed framework ProgAttn-DAG.

by DialogueRNN for temporal contextual modeling, resulting in 256-dimensional contextual audio representations, denoted as y_i^a.

D. Multimodal Fusion

Inspired by the progressive attention mechanism [20] and the bidirectional multi-head cross-attention in MultiEMO [21], we propose an enhanced progressive attention-based multimodal fusion network (ProgAttn).

Each modality layer has three steps: visual features (Query) attend to text, the fused result attends to audio, then a feed-forward network refines it, reducing computation and inter-modal interference. Given the textual features $y^t = [y_1^t, \ldots, y_n^t]^T$ and audio features $y^a = [y_1^a, \ldots, y_n^a]^T$, the visual feature Queries from layer $l-1$ are represented as $F^{v^{(l-1)}} = [f_1^{v^{(l-1)}}, \ldots, f_n^{v^{(l-1)}}]^T$.

In the first stage of ProgAttn, visual features from layer $(l-1)$ serve as Query, and textual features as Key and Value in a multi-head attention architecture. After projecting features to Q, K, and V, we apply scaled dot-product cross-modal attention and concatenate the heads. The result is integrated with the original visual features via residual connection and layer normalization to obtain $F^{vt^{(l)}}$. The computation is formulated as follows in (1)–(4):

$$[Q_h^{vt^{(l)}}, K_h^{vt^{(l)}}, V_h^{vt^{(l)}}] = [F^{v^{(l-1)}} W^{Q_h^{vt^{(l)}}}, y^t W^{K_h^{vt^{(l)}}}, y^t W^{V_h^{vt^{(l)}}}] \tag{1}$$

$$A_h^{vt^{(l)}} = \text{Softmax}\left(\frac{Q_h^{vt^{(l)}} K_h^{vt^{(l)T}}}{\sqrt{d_{K_h^{vt^{(l)}}}}}\right) \cdot V_h^{vt^{(l)}} \tag{2}$$

$$MH^{vt^{(l)}} = \text{Concat}(A_1^{vt^{(l)}}, \ldots, A_H^{vt^{(l)}}) \cdot W^{O^{vt^{(l)}}} \tag{3}$$

$$F^{vt^{(l)}} = \text{LayerNorm}(F^{v^{(l-1)}} + MH^{vt^{(l)}}) \tag{4}$$

where $W^{Q_h^{vt^{(l)}}}$, $W^{K_h^{vt^{(l)}}}$, $W^{V_h^{vt^{(l)}}}$, and $W^{O^{vt^{(l)}}}$ are learnable projection matrices for the h-th attention head; H is the number of attention heads; and Concat$(\cdot)$ denotes concatenation.

In the second stage, the text-enhanced visual features serve as the new Q, and audio features as K and V. A similar multi-head attention process is applied to obtain the final representation $F^{vta^{(l)}}$, which is the output of $ProgAttn^{vta}$. The computation is formulated as follows in (5)–(8):

$$[Q_h^{vta^{(l)}}, K_h^{vta^{(l)}}, V_h^{vta^{(l)}}] = [F^{vt^{(l)}} W_h^{Q^{vta^{(l)}}}, y^a W_h^{K^{vta^{(l)}}}, y^a W_h^{V^{vta^{(l)}}}] \tag{5}$$

$$A_h^{vta^{(l)}} = \text{Softmax}\left(\frac{Q_h^{vta^{(l)}} K_h^{vta^{(l)T}}}{\sqrt{d_{K_h^{vta^{(l)}}}}}\right) \cdot V_h^{vta^{(l)}} \tag{6}$$

$$MH^{vta^{(l)}} = \text{Concat}\left(A_1^{vta^{(l)}}, \ldots, A_H^{vta^{(l)}}\right) \cdot W^{O^{vta^{(l)}}} \tag{7}$$

$$F^{vta^{(l)}} = \text{LayerNorm}\left(F^{vt^{(l)}} + MH^{vta^{(l)}}\right) \tag{8}$$

where $W_h^{Q^{vta^{(l)}}}$, $W_h^{K^{vta^{(l)}}}$, $W_h^{V^{vta^{(l)}}}$, and $W^{O^{vta^{(l)}}}$ denote learnable projection matrices.

Finally, two feed-forward neural networks are applied to perform non-linear transformations on the cross-modal multi-head attention outputs, enhancing the model's representational capacity. $F^{V^{(l)}}$ constitutes the visual feature representation at layer l, formulated as follows in (9)–(11):

$$FFN_1^{v^{(l)}} = \max(0, F^{vta^{(l)}} W^{FFN_1^{v^{(l)}}} + b^{FFN_1^{v^{(l)}}}) \tag{9}$$

$$FFN_2^{v^{(l)}} = FFN_1^{v^{(l)}} \cdot W^{FFN_2^{v^{(l)}}} + b^{FFN_2^{v^{(l)}}} \tag{10}$$

$$F^{V^{(l)}} = \text{LayerNorm}(F^{vta^{(l)}} + FFN_2^{v^{(l)}}) \tag{11}$$

where $b^{FFN_1^{v^{(l)}}}$ and $b^{FFN_2^{v^{(l)}}}$ are bias terms.

Through stacking T layers of ProgAttn, feature fusion is progressively deepened. Each layer's output serves as the Query for the next, enabling progressive feature enhancement. After multi-layer attention, the visual, textual, and audio representations are concatenated to form a multimodal feature vector for each utterance, which is then input to the graph neural network.

Therefore, for an utterance u_i, its corresponding multimodal feature vector can be represented as:

$$F^{VTA^{(l)}} = F^{V^{(l)}} \oplus F^{T^{(l)}} \oplus F^{A^{(l)}} \tag{12}$$

where $\oplus$ denotes vector concatenation operation, and $F^{VTA^{(l)}}$ serves as the input to the subsequent DAG network, providing the foundation for fine-grained inter-modal interaction.

E. Multi-layer DAG Networks

Directed Acyclic Graphs (DAGs) with directionality and acyclicity suit conversational emotion settings with unpredictable future utterances. Inspired by DAGNN [22], we implement a DAG-based network that models multimodal interaction paths while removing computational redundancy and cyclic dependencies. Nodes denote utterances; directed edges carry information from historical nodes to the current node. Each layer performs cross-modal information aggregation and gated feature updating.

Cross-modal Information Aggregation: We employ a dynamic attention mechanism to aggregate contextual information from historical nodes, adaptively adjusting inter-modal interaction weights. When visual and acoustic signals contradict textual content, the attention mechanism suppresses noisy modalities. The attention coefficient is calculated as:

$$\alpha_{ij}^{VTA^{(l)}} = \text{Softmax}\left(W_\alpha^l [F_j^{VTA^{(l)}} \| F_i^{VTA^{(l-1)}}]\right)\Big|_{j \in \mathcal{N}_i^{VTA}} \tag{13}$$

where $F_j^{VTA^{(l)}}$ represents the hidden state of candidate nodes in the neighborhood, and $F_i^{VTA^{(l-1)}}$ denotes the hidden state of the current node from the previous layer.

Gated Feature Updating: Neighborhood information aggregation is achieved by computing the weighted sum of adjacent node features:

$$M_i^{VTA^{(l)}} = \sum_{j \in \mathcal{N}_i^{VTA}} \alpha_{ij}^{VTA^{(l)}} W_{s_{ij}}^l F_j^{VTA^{(l)}} \tag{14}$$

where s_{ij} is the edge attribute, $s_{ij} = 1$ for same-speaker utterances, $s_{ij} = 0$ for different speakers. The output $M_i^{VTA^{(l)}}$ captures contextually integrated information from neighboring nodes.

For gated feature updating, we use a GRU-based unit that processes the previous hidden state and aggregated neighborhood information to update node features, refining local information. The computation is formulated as:

$$\tilde{F}_i^{VTA^{(l)}} = \text{GRU}_F^l(F_i^{VTA^{(l-1)}}, M_i^{VTA^{(l)}}) \tag{15}$$

where $\tilde{F}_i^{VTA^{(l)}}$ denotes the updated hidden state. Subsequently, contextual information updating is performed through:

$$C_i^{VTA^{(l)}} = \text{GRU}_M^l(M_i^{VTA^{(l)}}, F_i^{VTA^{(l-1)}}) \tag{16}$$

This bidirectional interaction is achieved by inverting the GRU sequence, modeling historical context influence on the current node. Finally, the comprehensive multimodal representation is obtained by integrating information from multiple graph layers:

$$F_i^{VTA^{(l)}} = \bigoplus_{l=0}^{L} (\tilde{F}_i^{VTA^{(l)}} + C_i^{VTA^{(l)}}) \tag{17}$$

F. Emotion Classification

The final output $F_i^{VTA^{(l)}}$ from the DAG network is projected by a two-layer Feed-Forward Network (FFN) to logits and then to probabilities and labels:

$$Z_i = W_2 \cdot \text{ReLU}(W_1 \cdot F_i^{VTA^{(l)}} + b_1) + b_2 \tag{18}$$

$$P(e|u_i) = \text{Softmax}(Z_i) \tag{19}$$

$$\hat{e}_i = \arg\max P(e|u_i) \tag{20}$$

where W_1, W_2, b_1, b_2 are learnable parameters, $\hat{e}_i$ is the final predicted emotion label for utterance u_i.

G. CL Strategies

Curriculum learning comprises two components: the Difficulty Measure Function (DMF) and the Training Scheduler. The DMF quantifies dialogue complexity based on emotional transition frequency, defined as a shift in emotional expression by the same speaker across consecutive utterances. The formula is:

$$\text{DMF}(c_i) = \frac{N_{shift}(c_i) + N_{sp}(c_i)}{N_u(c_i) + N_{sp}(c_i)} \tag{21}$$

where $N_{shift}(c_i)$ denotes the number of emotional transitions in dialogue c_i, $N_u(c_i)$ represents the total number of utterances, and $N_{sp}(c_i)$ indicates the number of speakers (serving as a smoothing factor to mitigate bias in short dialogues).

Dialogues are sorted by DMF and split into k difficulty buckets; training starts with D_1 and at fixed intervals adds (D_2, D_3, ...) until all are merged for joint optimization. This curriculum yields phased difficulty growth, improving learning efficiency and alleviating data imbalance.

H. Training Objectives

We optimize the proposed framework by minimizing three complementary loss functions. The total loss is:

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{swfc}\mathcal{L}_{swfc} + \lambda_{hgr}\mathcal{L}_{hgr} \tag{22}$$

where λ_{ce}, λ_{swfc}, and λ_{hgr} are scalar weighting hyperparameters.

Cross-entropy (CE) loss ($\mathcal{L}_{ce}$) [23] measures the discrepancy between predicted and true emotion labels:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{23}$$

where N is batch size, C is the number of emotion classes, $y_{i,c} \in \{0,1\}$ is the one-hot ground-truth indicator for sample i and class c, and $p_{i,c} = P(e = c \,|\, u_i)$ is the predicted softmax probability.

To address class imbalance, we use sample-weighted focal contrastive (SWFC) loss [24] ($\mathcal{L}_{swfc}$):

$$q_{i,j} = \frac{\exp(s_{i,j}/\tau)}{\sum_{k\in\mathcal{J}_i}\exp(s_{i,k}/\tau)}, \quad j \in \mathcal{J}_i \tag{24}$$

$$q_i^+ = \sum_{j\in\mathcal{P}_i} q_{i,j} \tag{25}$$

$$\mathcal{L}_{swfc} = -\frac{1}{N}\sum_{i=1}^{N} w_i(1-q_i^+)^{\gamma}\log(q_i^+) \tag{26}$$

where $s_{i,j}$ is the similarity between embeddings z_i and z_j; τ is the temperature; $\mathcal{J}_i = \{k \mid k \neq i\}$ is the non-anchor index set; $\mathcal{P}_i = \{j \in \mathcal{J}_i \mid y_j = y_i\}$ is the positive set (negatives $\mathcal{N}_i = \mathcal{J}_i \setminus \mathcal{P}_i$); $q_{i,j}$ is the normalized contrastive probability; q_i^+ is the aggregated positive probability; w_i is a (class/sample) weight; γ is the focal focusing factor.

To enhance cross-modal feature correlation, we employ Soft-HGR loss [25] ($\mathcal{L}_{hgr}$). Let $\mathcal{M}$ be the modality set and $\mathcal{P}_{mod} = \{(m,n) \mid m < n,\ m,n \in \mathcal{M}\}$. For each sample i and modality m, raw features $h_i^{(m)}$ are projected to a shared d-dimensional space and centered. The Soft-HGR term is:

$$g_i^{(m)} = h_i^{(m)} - \frac{1}{N}\sum_{r=1}^{N} h_r^{(m)} \tag{27}$$

$$\mathcal{L}_{hgr} = -\frac{1}{|\mathcal{P}_{mod}|}\sum_{(m,n)\in\mathcal{P}_{mod}} \Big(\mathbb{E}_i\big[(g_i^{(m)})^{\top} g_i^{(n)}\big] - \tfrac{1}{2}\,\mathrm{Tr}\left(\Sigma^{(m)}\Sigma^{(n)}\right)\Big). \tag{28}$$

where $\Sigma^{(m)} = \mathbb{E}_i\big[(g_i^{(m)} - \mathbb{E}_i[g_i^{(m)}])(g_i^{(m)} - \mathbb{E}_i[g_i^{(m)}])^{\top}\big]$; $\mathbb{E}_i[\cdot]$ denotes the empirical batch expectation. Minimizing $\mathcal{L}_{hgr}$ maximizes cross-modal correlation while discouraging redundant second-order structure.

TABLE I
DATASET STATISTICS

Dataset	Convs.			Utts.			Avg.
	Train	Val	Test	Train	Val	Test	utt.
IEMOCAP	120	31	31	5810	1623	1623	66.8
MELD	1038	114	280	9989	1109	2610	9.57

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We validate the performance of the ProgAttn-DAG model on two benchmark datasets widely used for emotion recognition. The data distribution of these datasets is shown in Table I.

IEMOCAP [26] is an approximately 12-hour multimodal dataset of dyadic dialogues among 10 speakers (pairs) with six emotions: happiness, sadness, neutral, anger, excitement, frustration.

MELD [27] is a multimodal multi-party dialogue dataset from the TV series "Friends," with 13,709 video clips and speakers; each utterance is annotated with one of seven emotions: neutral, surprise, fear, sadness, joy, disgust, angry.

Evaluation Metrics: This study employs overall recognition accuracy (Acc) and weighted F1-score (W-F1) as standard evaluation metrics to assess the emotional recognition performance of the model. These quantitative indicators comprehensively reflect the model's effectiveness in multi-class emotion recognition tasks.

B. Baselines

- **Sequence-based Methods**, DialogueRNN [3], MMDFN [28], MVN [29];
- **Graph-based Methods**, DialogueGCN [4], MMGCN [8], GraphMFT [30], DAG-ERC [6], MultiDAG+CL [31].

C. Implementation Details

Hyperparameter optimization is performed separately for each dataset using hold-out validation with dedicated validation sets. For IEMOCAP, we use a batch size of 32, learning rate 4.5e-5, weight decay 1e-4, 8 layers of 512 dimensions, 8 attention heads, and loss weights of 0.45 (SWFC), 0.35 (HGR), and 0.2 (CE), with 15 curriculum steps. For MELD, the network architecture remains the same, with learning rate 5e-6, weight decay 1e-3, dropout 0.45, loss weights 0.55 (SWFC), 0.15 (HGR), and 0.3 (CE), and 20 curriculum steps.

V. RESULTS AND ANALYSIS

A. Comparison with Baseline Models

Table II compares ProgAttn-DAG with baseline models on IEMOCAP and MELD. The experimental results demonstrate that ProgAttn-DAG achieves superior weighted F1-score and accuracy. Compared with MultiDAG+CL on IEMOCAP, ProgAttn-DAG improves weighted F1-score and accuracy by 1.96% and 1.75%, respectively. On the MELD dataset, relative to the MVN model, our model achieves a 20.11%

TABLE II

PERFORMANCE COMPARISON OF PROGATTN-DAG AND BASELINE MODELS ON IEMOCAP AND MELD DATASETS. $M*$ RESULTS REFER TO THE EXPERIMENTAL RESULTS RETRAINED ON THE 3090 SERVER BY US, WHILE OTHERS ARE FROM THE ORIGINAL PAPERS. BOLD FONT DENOTES THE HIGHEST PERFORMANCE.

Model	IEMOCAP								MELD								
	Emotion Categories (F1)						Overall		Emotion Categories (F1)							Overall	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc.(%)	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry	Acc.(%)	W-F1
DialogueRNN*	33.19	80.78	56.36	61.64	71.27	58.43	62.54	61.78	75.56	48.03	0.00	26.48	52.68	0.00	45.11	59.23	57.72
DialogueGCN*	31.58	74.40	49.93	60.59	64.89	60.96	59.03	58.46	76.76	25.65	0.00	5.29	43.36	0.00	37.76	55.71	51.79
MMGCN	42.34	78.67	61.73	69.00	74.33	62.32	—	66.22	—	—	—	—	—	—	—	—	58.65
MMDFN	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18	77.76	50.69	—	22.93	54.78	—	47.82	62.49	59.46
GraphMFT	45.99	**83.12**	63.08	70.30	**76.92**	63.84	67.90	68.07	—	—	—	—	—	—	—	61.30	58.37
DAG-ERC	—	—	—	—	—	—	—	68.03	—	—	—	—	—	—	—	—	63.65
MVN	55.75	73.30	61.88	65.96	69.50	64.21	65.32	65.44	76.65	53.18	11.70	21.82	53.62	21.86	42.55	61.29	59.03
MultiDAG+CL*	**68.17**	69.06	69.32	**82.11**	43.48	**69.68**	68.68	68.69	77.36	**58.48**	26.15	37.67	61.21	28.87	49.45	64.18	63.74
ProgAttn-DAG (Ours)*	63.06	80.85	**70.08**	66.50	76.22	64.99	**70.43**	**70.65**	**78.83**	58.25	**27.91**	**41.93**	**63.73**	**32.43**	**50.91**	**66.13**	**65.47**

TABLE III

ABLATION STUDY OF PROGATTN-DAG COMPONENTS.

Method	IEMOCAP		MELD	
	Acc.(%)	W-F1	Acc.(%)	W-F1
T+A	64.63	64.89	65.95	65.15
T+V	64.57	60.86	64.10	63.72
w/o ProgAttn	69.13	69.22	65.59	64.67
w/o DAG	69.01	69.35	65.06	64.78
w/o CL	69.38	69.58	65.75	65.13
Ours	**70.43**	**70.65**	**66.13**	**65.47**

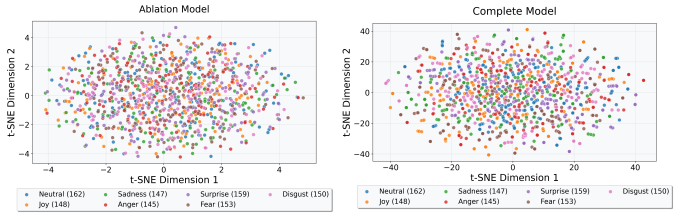

Fig. 3. Visualization of model's outputs using t-SNE with and without ProgAttn.

improvement in the F1-score for recognizing the minority emotion category Sadness. By leveraging ProgAttn to model conversational context and curriculum learning to address data imbalance, ProgAttn-DAG enables effective cross-modal interaction, fine-grained feature fusion, and improved recognition of minority and semantically similar emotion categories.

B. Ablation Study

To further evaluate the effectiveness of the proposed ProgAttn-DAG framework, we conduct ablation experiments on the IEMOCAP and MELD datasets by removing one module at a time. The results are summarized in Table III.

Effect of Different Modalities: The ablation results demonstrate that multimodal inputs outperform unimodal inputs significantly. It underscores the importance of multimodal fusion in capturing complementary information across modalities.

Effect of Progressive Attention Mechanism: To qualitatively analyze ProgAttn's effectiveness, we visualized feature distributions using t-SNE [32] on the MELD dataset. As shown in Fig. 3, without ProgAttn, the feature representations exhibit confusion and overlap across different emotion categories. The ablation model (left panel) shows that emotional features are widely dispersed with poor class separation, indicating that the baseline fusion strategy fails to capture discriminative multimodal patterns effectively. With ProgAttn, the feature distribution demonstrates remarkable improvement in clustering quality and class separability. The complete model (right panel) shows that emotional categories form distinct, well-separated clusters with significantly reduced inter-class overlap, indicating that ProgAttn successfully captures complementary information across text, audio, and visual modalities.

Effect of Multi-layer DAG Networks: The multi-layer DAG networks explicitly model dependencies between utterances and modalities, facilitating fine-grained contextual modeling. As shown in Table III, removing DAG leads to lower accuracy (down 1.42% on IEMOCAP and 1.07% on MELD), demonstrating the effectiveness of DAG-based contextual modeling.

Effect of Curriculum Learning: CL facilitates sample-difficulty-based training optimization. Ablation results show that removing CL degrades model accuracy on minority emotions, validating effectiveness in addressing data imbalance.

VI. CONCLUSION

This paper presents ProgAttn-DAG, a framework for multimodal conversational emotion recognition that hierarchically integrates text, audio, and visual modalities, addressing modality heterogeneity and enhancing cross-modal representations. ProgAttn provides fine-grained adaptive fusion, multi-layer DAG networks model utterance–modality dependencies, and Curriculum Learning (CL) progressively adjusts sample difficulty to mitigate data imbalance and improve minority emotion recognition. Experiments on benchmark datasets show superior performance over existing methods, validating its attention-based fusion, DAG modeling, and CL strategies.

REFERENCES

[1] M. Minsky, *The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind*. Simon & Schuster, 2007.

[2] S. Planalp, J. Fitness, and B. A. Fehr, "The roles of emotion in relationships," in *Cambridge Handbooks in Psychology*, 2nd ed. Cambridge University Press, 2018, pp. 256–268.

[3] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: an attentive RNN for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 6818–6825.

[4] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: a graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 154–164.

[5] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 5415–5421.

[6] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 1551–1560.

[7] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: COntextualized GNN based multimodal emotion recognitioN," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4148–4164. [Online]. Available: https://aclanthology.org/2022.naacl-main.306/

[8] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 5666–5675. [Online]. Available: https://aclanthology.org/2021.acl-long.440/

[9] Z. Han, B. Zhu, Y. Xu, P. Song, and X. Yang, "Benchmarking and bridging emotion conflicts for multimodal emotion reasoning," in *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. Dublin, Ireland: Association for Computing Machinery, October 2025. [Online]. Available: https://doi.org/10.1145/3746027.3754856

[10] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th International Conference on Machine Learning*, ser. ICML '09. Montreal, Canada: ACM, 2009, pp. 41–48.

[11] Y. Mao, X. Li, H. Zhang, J. Chen, and Y. Liu, "Hybrid curriculum learning for emotion recognition in conversation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1234–1241.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.

[13] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 2261–2269.

[14] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognition*, vol. 85, pp. 1–12, 2019.

[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[16] J. Lee and W. Lee, "CoMPM: context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 5669–5679.

[17] T. Kim and P. Vossen, "EmoBERTa: speaker-aware emotion recognition in conversation with RoBERTa," *arXiv preprint arXiv:2108.12009*, 2021. [Online]. Available: https://arxiv.org/abs/2108.12009

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: a robustly optimized BERT pretraining approach," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, arXiv:1907.11692. [Online]. Available: https://arxiv.org/abs/1907.11692

[19] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 1459–1462.

[20] W. Zhang, X. Li, and Y. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '18. Piscataway, NJ, USA: IEEE, 2018, pp. 5226–5235.

[21] T. Shi and S.-L. Huang, "MultiEMO: an attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations," *arXiv*, vol. 2311.15316, 2023, arXiv preprint arXiv:2311.15316. [Online]. Available: https://arxiv.org/abs/2311.15316

[22] V. Thost and J. Chen, "Directed acyclic graph neural networks," *arXiv*, 2021, arXiv preprint arXiv:2101.07965. [Online]. Available: https://arxiv.org/abs/2101.07965

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Lake Tahoe, NV, USA: Curran Associates, Inc., 2012, pp. 1097–1105.

[24] Y. Ding, Z. Li, D. Huang, Z. Li, and K. Zhang, "Enhancing multi-view stereo with contrastive matching and weighted focal loss," *arXiv preprint arXiv:2206.10360*, 2022. [Online]. Available: https://arxiv.org/abs/2206.10360

[25] Y. Wang, Y. Wang, Y. Yang, J. Sun, J. Hsiao, and D. Zhang, "Deep multimodal fusion by channel exchanging," in *Advances in Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc., 2019, pp. 4824–4834.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[27] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: a multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '19, vol. 1: Long Papers. Association for Computational Linguistics, 2019, pp. 527–536.

[28] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7037–7041.

[29] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang, "A multi-view network for real-time emotion recognition in conversations," *Knowledge-Based Systems*, vol. 236, p. 107751, 2022.

[30] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphMFT: a graph network based multimodal fusion technique for emotion recognition in conversation," *Neurocomputing*, vol. 550, p. 126427, 2023.

[31] C.-V. T. Nguyen, C.-B. Nguyen, Q.-T. Ha, and D.-T. Le, "Curriculum learning meets directed acyclic graph for multimodal emotion recognition," https://arxiv.org/abs/2402.17269, 2024, arXiv:2402.17269v2. [Online]. Available: https://github.com/vanntc711/MultiDAG-CL

[32] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, vol. 15. Cambridge, MA, USA: MIT Press, 2002, pp. 857–864.