BIG-FUSION: Brain-Inspired Global-Local Context Fusion Framework for Multimodal Emotion Recognition in Conversations

Yusong Wang[1,2,*], Xuanye Fang[3,*], Huifeng Yin[4,*]
Dongyuan Li[2], Guoqi Li[5,6], Qi Xu[3,†], Yi Xu[3], Shuai Zhong[1], Mingkun Xu[1,4,†]

[1]Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, China
[2]Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan
[3]School of Computer Science and Technology, Dalian University of Technology, Dalian, China
[4]Center for Brain Inspired Computing Research (CBICR), Department of Precision Instrument, Tsinghua University, Beijing, China
[5]Institute of Automation, Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Chinese Academy of Sciences, Beijing, China
[6]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract

Considering the importance of capturing both global conversational topics and local speaker dependencies for multimodal emotion recognition in conversations, current approaches first utilize sequence models like Transformer to extract global context information, then apply Graph Neural Networks to model local speaker dependencies for local context information extraction, coupled with Graph Contrastive Learning (GCL) to enhance node representation learning. However, this sequential design introduces potential biases: the extracted global context information inevitably influences subsequent processing, compromising the independence and diversity of the original local features; current graph augmentation methods in GCL cannot consider both global and local context information in conversations to evaluate the node importance, hindering the learning of key information. Inspired by the human brain excels at handling complex tasks by efficiently integrating local and global information processing mechanisms, we propose an aligned global-local context fusion framework for sequence-based design to address these problems. This design includes a dual-attention Transformer and a dual-evaluation method for graph augmentation in GCL. The dual-attention Transformer combines global attention for overall context extraction with sliding-window attention for local context capture. Meanwhile, spiking neuron dynamics are introduced to enhance the representation capability of the extracted features, supporting the effective the interaction between different modalities from these global and local cues. The dual-evaluation method in GCL comprises global importance evaluation to identify nodes crucial for overall conversation context, and local importance evaluation to detect nodes significant for local semantics, generating augmented graph views that preserve both global and local information. This approach ensures balanced information processing throughout the pipeline, enhancing biological plausibility and achieving superior emotion recognition.

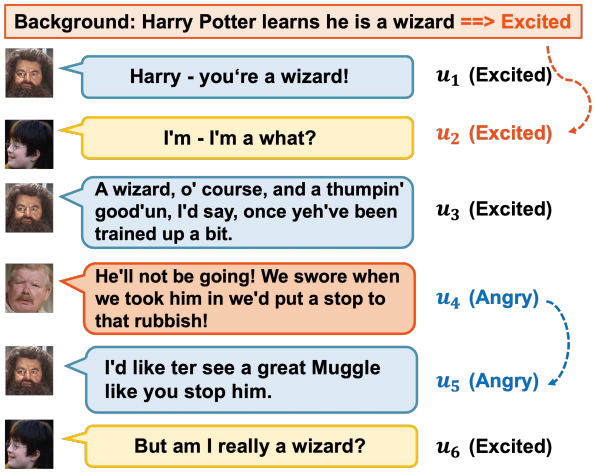

Figure 1: An example of the conversation from *Harry Potter*, showcasing the interplay between global and local context.

Introduction

Multimodal emotion recognition in conversations (MERC) identifies the emotion state of each utterance in the conversion with visual, audio, and text information (Zhang et al. 2024). It has received increasing attention due to its different applications like conversational speech synthesis (Liu et al. 2024) and social media analytics (Poria et al. 2019).

As shown in Figure 1, two main factors influence the expression of emotions in a conversation: 1) Global Contextual Information. The emotion states expressed in utterances are directly correlated with the overall conversation context. In Figure 1, the topic "Harry Potter learns he is a wizard" with its positive tone results in the majority of sentences expressing positive emotions in the conversation. For instance, the utterance u_2 is more likely to indicate excitement when considering the main context of the conversation. 2) Local Context Information. The emotion of the current utterance can be

*These authors contributed equally.
†Corresponding author: xuqi@dlut.edu.cn,xumingkun@gdiist.cn

influenced by both the nearby speaker's utterances and those of others (e.g., u_5 is affected by the anger in u_4 although the holistic conversation tone is positive). Considering these factors, mainstream MERC methods employ a sequence model like Transformer (Vaswani et al. 2017) to capture the conversation's global context information, and then use Graph Neural Networks (GNNs) (Scarselli et al. 2009) to model the local utterances/speaker dependencies (Ghosal et al. 2019; Sun, Yu, and Fu 2021). Sequence models mainly utilize the attention mechanism across utterances to enhance the flow of global context information and generate a representation for each utterance. Graph models capture the dependencies between utterances/speakers and aggregate information from surrounding contextual utterances to update the representation of the current utterance, enabling an understanding of the conversation's emotional dynamics. By integrating both components to capture both global and local context information, these MERC methods achieve remarkable accuracy (Joshi et al. 2022). The latest, related method introduces Graph Contrastive Learning (GCL) to enhance the node representation and alleviate the over-smoothing problem, further improving performance (Li et al. 2023b).

Although this design has achieved great success, it still faces challenges. It employs a sequential design: a sequence model captures global context, followed by a graph model for local context. This design introduces bias, as the global context influences the subsequent local context extraction, resulting in a loss of original independence and diversity. Besides, GCL-based methods often rely on pre-defined view generation methods, such as random node/edge masking (You et al. 2020), or automated graph augmentation methods that select nodes by global semantics change (Yin et al. 2021; Wei et al. 2023). These approaches overlook the dual influence on both global and local contexts for MERC, leading to suboptimal node/edge selection and compromising the learning of critical information and patterns of the graph.

On the contrary, the human brain excels at handling complex tasks like the MERC task by efficiently integrating both local and global information processing mechanisms at every stage, (Wu et al. 2020; Wang et al. 2021; Jiang et al. 2024). which is a crucial element missing in current sequential structures and GCL. To bridge this gap, we propose a **b**rain-**i**nspired **g**lobal-local context **fusion** framework, called BIG-FUSION. To capture global context information while preserving the local one, we propose a dual-attention mechanism for Transformer, which combines global attention to capture overarching context information with sliding-window attention to extract information from neighboring utterances. Building upon this dual-attention design, we incorporate spiking dynamics, which shows effectively simulate interactions between different modalities (Wysoski, Benusková, and Kasabov 2010), potentially enhancing the integration of multimodal information in emotion recognition (Wang et al. 2024). We utilize rate coding in spiking neural networks (SNNs) and spike frequencies to interpret and decode attention outputs (Xu et al. 2021b). Spiking dynamics, by mimicking the functioning of human neurons, further reinforce our brain-inspired design. To benefit GCL learning comprehensive information, we introduce a dual-evaluation approach aimed at node selection for automated graph augmentation to create augmented graph views that preserve both global and local context information: for global importance, we perturb individual nodes and compute the mutual information between the original and perturbed graphs; for local importance, we perturb a node within a defined subgraph and compute the mutual information between the original and perturbed sub-graphs. The main contributions are:

• We propose BIG-FUSION, a brain-inspired MERC model that incorporates global-local information processing mechanisms and spiking dynamics to perform the MERC task in a biologically plausible manner.

• We propose a dual-attention for Transformer and a dual-evaluation method for GCL to enable simultaneous processing of both global/local context information, addressing the information extraction issues in current sequential designs.

• Extensive experiments on two representative MERC benchmarks show the state-of-the-art (SOTA) performance.

Related Work

Multimodal Emotion Recognition in Conversation

MERC methods are categorized two main ways: sequence-based approaches use RNNs or Transformers for global context modeling (Majumder et al. 2018; Yang et al. 2023), while graph-based approaches leverage GNNs for local context modeling and speaker dependencies (Zhang, Chen, and Chen 2023; Tu et al. 2024). Current MERC methods have started to model both global and local context information by combining sequence-based and graph-based models, achieving complete context modeling for conversations and advanced performance (Joshi et al. 2022; Li et al. 2023b; Tu et al. 2024). However, they typically employ a sequential design, using a sequence model to first capture global context information, followed by a graph model to extract local context information. In this case, global context information interferes the local context information extraction and the downstream task relies on local context information, compromising the integrity of the information. Thus, we design an aligned global-local context fusion framework for both components to maintain an information balance.

Graph Contrastive Learning

Graph methods face over-smoothing issues (Tan et al. 2024; Li et al. 2023a), hindering emotion representation. Li et al. (2023b) introduced GCL for better node distinction. However, they use conventional pre-defined view generation techniques such as random dropout/masking to augment different graph views, ignoring the discrepancy in the impact of different nodes/edges. Current GCL methods assess importance via global semantics, overlooking nodes key to local-global semantics (Suresh et al. 2024; Yin et al. 2021; Wei et al. 2023). In the MERC task, these methods may hinder the learning of comprehensive context information and patterns within graphs. Since GCL mainly focuses on the relative relationships between samples rather than their absolute characteristics, it tends to be invariant to corruption induced by such graph augmentation scheme (Xiao et al. 2020). To overcome this problem, we propose a dual-evaluation

method that can evaluate the node importance on both local and global semantics.

Global-Local Information in the Human Brain

The human brain employs both global and local processing mechanisms to achieve high accuracy in complex tasks (Wu et al. 2020). During global context processing, the human brain preserves relevant local information, utilizing it to enrich and refine the overall contextual understanding. When processing local information, the human brain considers global context to guide and modulate responses, thereby enhancing decision accuracy (Lin et al. 2019; Primativo and Arduino 2023). The human brain's integration of global and local information processing inspires our model design.

Spiking Attention

SNNs model the neural dynamics of biological neurons through discrete spike discharges, achieving higher biological interpretability (Xu et al. 2023; Yin et al. 2024; Xu et al. 2021a; Zhao et al. 2022). Spiking attention, closely mimicking human brain's functioning, dynamically adjusts neuron activation thresholds based on information context, facilitating multimodal information capture (Wang et al. 2024).

Preliminary

Notations and Task Definition

In the MERC, a training dataset $\mathcal{D} = \{(\mathcal{C}_i, \mathcal{Y}_i)\}_{i=1}^{N}$ is given, where $\mathcal{C}_i$ represents the i-th conversation, each conversation contains utterances $\mathcal{C}_i = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$, and $\mathcal{Y}_i \in \mathbf{Y}^n$, given label set $\mathbf{Y} = \{y_1, \ldots, y_l\}$ of l emotion classes. Let $\mathbf{X}^v$, $\mathbf{X}^a$, $\mathbf{X}^t$ be the visual, audio, and text feature spaces, respectively. MERC aims to learn a function $\mathbf{F} : \mathbf{X}^v \times \mathbf{X}^a \times \mathbf{X}^t \rightarrow \mathbf{Y}$ that can recognize the emotion label for each utterance.

Utterance-level Encoder

Here, we introduce the feature extraction of each modality: **Audio modality.** We use OpenSmile (Eyben, Wöllmer, and Schuller 2010) to extract acoustic features. Then, a fully connected layer is used to reduce the dimension of acoustic feature representations to 1,582 for the IEMOCAP dataset and 300 for the MELD dataset. **Text modality.** We employ RoBERTa model (Liu et al. 2019) to extract textual features and the dimension is 1024 for both datasets. **Visual modality.** We use a pretrained DenseNet (Huang, Liu, and Weinberger 2016) and the dimension is 342 for both datasets.

Graph Construction

Graph construction aims to establish relationships between utterances, preserving both intra-speaker and inter-speaker dependencies in a conversation. We define the i-th conversation with P speakers as $\mathcal{C}_i = \{\mathcal{U}^{S_1}, \ldots, \mathcal{U}^{S_P}\}$, where $\mathcal{U}^{S_i} = \{\boldsymbol{u}_1^{S_i}, \ldots, \boldsymbol{u}_m^{S_i}\}$ represents the set of utterances spoken by speaker S_i. In the graph, nodes represent utterances, and directed edges represent speaker relations: $\mathcal{R}_{ij} = \boldsymbol{u}_i \rightarrow \boldsymbol{u}_j$, with the arrow indicating the speaking order. $\mathcal{R}_{intra} \in \mathcal{U}^{S_i} \rightarrow \mathcal{U}^{S_i}$ captures the intra-relations between utterances from same speaker, while $\mathcal{R}_{inter} \in \mathcal{U}^{S_i} \rightarrow \mathcal{U}^{S_j}, i \neq j$ captures the inter-relations between utterances from different speakers. Node representations are initialized using the output from the Transformer, and the neighborhood range for each node is determined by a specified window size w.

Graph Contrastive Learning

GCL utilizes a graph encoder like Relational Graph Convolutional Network (RGCN) (Schlichtkrull et al. 2017) to extract node hidden representations. For original and augmented graphs, the hidden representations are both denoted as $\mathbf{H} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\mathbf{H}^{'} = \{\boldsymbol{x}_1^{'}, \ldots, \boldsymbol{x}_n^{'}\}$, where $\boldsymbol{x}_i$ represents the hidden representation of the i-th node. GCL encompasses both intra-view and inter-view mechanisms to learn distinctive node representations. Given the definition of positive and negative pairs as $(\boldsymbol{x}_i, \boldsymbol{x}_i^{'})^{+}$ and $(\boldsymbol{x}_i, \boldsymbol{x}_j^{'})^{-}$, where $i \neq j$, the inter-view loss for the i-th node is:

$$\mathcal{L}_{inter}^{i} = -\log \frac{\exp(\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_i^{'}))}{\sum_{j=1}^{n} \exp(\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^{'}))}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity. Intra-view GCL regards all nodes in the original view except the anchor node as negatives. Given negative samples pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)^{-}$ where $i \neq j$, the intra-view loss for the i-th node is:

$$\mathcal{L}_{intra}^{i} = -\log \frac{\exp(\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_i^{'}))}{\sum_{j=1}^{n} \exp(\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j))}. \tag{2}$$

The contrastive objective function $\mathcal{L}_{ct}$ is formulated as:

$$\mathcal{L}_{cl} = \frac{1}{2n} \sum_{i=1}^{n} (\mathcal{L}_{inter}^{i} + \mathcal{L}_{intra}^{i}). \tag{3}$$

Methodology

In this section, we will introduce each part of BIG-FUSION sequentially and the overview is displayed in Figure 3.

Dual Attention

Dual-attention has global and local attention mechanisms to capture and preserve both key global and local context information. Let $\mathbf{X} \in \mathbb{R}^{B \times N \times D}$ be the input tensor, where B is the batch size, N is the sequence length, and D is the embedding dimension. Global Attention can be defined as:

$$\mathbf{A}_G(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B \times N \times d_k}$ are the query, key, and value respectively. d_k is the dimension of the key vectors. Local Attention employs a sliding-window to limit attention range and shift attention, and it is defined as:

$$\mathbf{A}_L(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Aggregate}[\text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right)\mathbf{V}_i] \tag{5}$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{B \times w \times d_k}$ are the i-th window of size w. The step size of the slide is 1 for capturing the local attention of each utterance. Aggregate represents weight-based fusion for windows. The final dual-attention mechanism is:

$$\mathbf{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \alpha\mathbf{A}_G(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + (1 - \alpha)\mathbf{A}_L(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \tag{6}$$

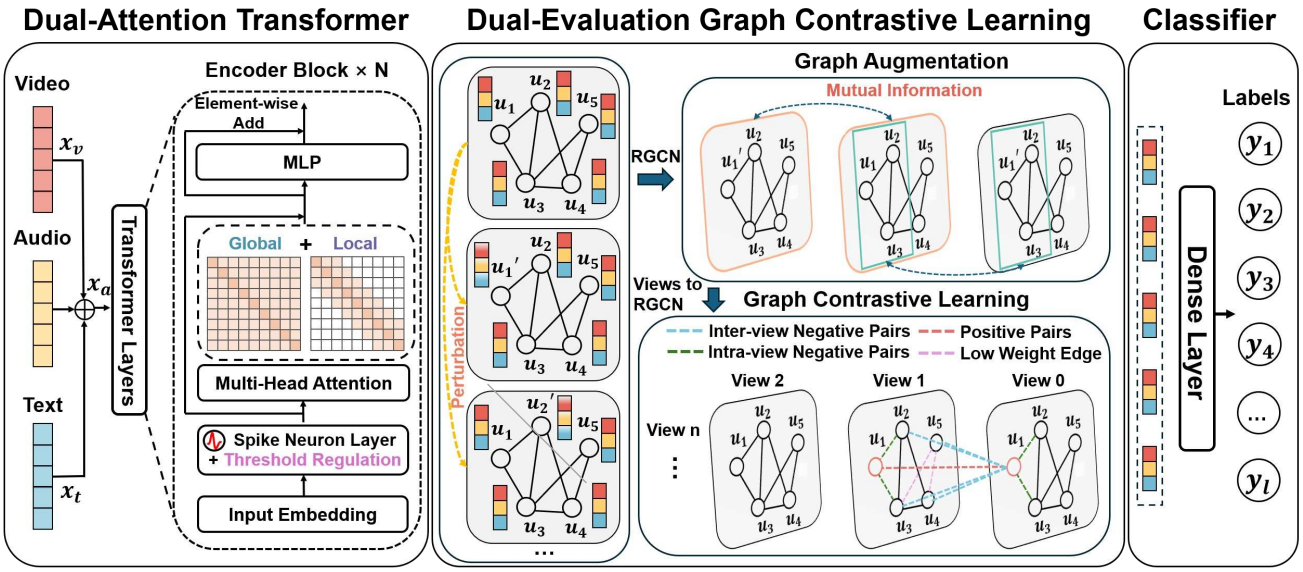

Figure 2: The BIG-FUSION comprises three key components: a dual-attention-based Transformer, a dual-evaluation-based graph augmentation for GCL, and a classifier.

where $\alpha \in [0, 1]$ is the weight coefficient to control the local and global information.

Spiking Attention

After the initial integration of multimodal stimuli, the information undergoes processing and refinement in the thalamus. Refinement specifically operates by regulating the spiking activity between neurons, simulating interactions between different modalities to handle high redundancy and complex structure multimodal data (Yu et al. 2023). Thus, we employ a spiking attention mechanism to enhance feature extraction capabilities and improve the model's biological interpretability:

$$\hat{\mathbf{Q}} = SN(\mathbf{Q}), \hat{\mathbf{K}} = SN(\mathbf{K}), \hat{\mathbf{V}} = SN(\mathbf{V}) \quad (7)$$

SN represents a spiking neuron layer and is formulated as:

$$H(t) = \beta U(t-1) \cdot (1 - \Theta(U(t-1) - V_{th})) + X(t) \quad (8)$$

where $H(t)$ and $U(t-1)$ denotes the current and last membrane potentials of the Leaky Integrate-and-Fire neuron, and β is the decay constant. $X(t)$ is the input excitation and Θ is the Heaviside function. We exploit rate coding in SNNs, leveraging spike frequencies to decode attention outputs. Considering that batch-processed inputs contain zero padding, we propose a threshold regulation mechanism. We continuously improve the threshold V_{th} as time steps increase to inhibit pulse firing of the input padding part. This adaptive threshold approach enables the model to effectively distinguish between meaningful input signals and padding. $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ in Eqs. 4, 5, 6 are substituted by those in Eq. 7.

Dual Node Importance Evaluation

To obtain an augmented graph $\mathbf{H}^{'}$ while keeping comprehensive information and patterns, graph augmentation should consider both global and local context information.

Global importance evaluation. The core idea is to identify the critical node $\boldsymbol{x}_i$ in $\mathbf{H}$ that influences the topic/background context information of the conversation and maximizes the mutual information between $\mathbf{H}^{'}$. Formally speaking, we define the graph representation as:

$$\mathbf{h_H} = \text{Aggregation}\left(\boldsymbol{x}_i \mid i \in V\right),$$

where V is the number of nodes in the graph and Aggregation refers to the mean aggregation of node features following the RGCN. We introduce Gaussian noise as perturbations to the initial representation of a node in a graph $\mathbf{H}$, i.e., $\boldsymbol{x}_i + \epsilon$, for $i \in V$, will result difference in the mutual information $I_G(\mathbf{h}^{'}_{\mathbf{H}}, \mathbf{h_H})$ and the formulation is defined as:

$$I_G(\mathbf{h}^{'}_{\mathbf{H}}, \mathbf{h_H}) = E(\mathbf{h}^{'}_{\mathbf{H}}) + E(\mathbf{h_H}) - E(\mathbf{h}^{'}_{\mathbf{H}}, \mathbf{h_H})$$

where E is the entropy. If ΔI_G is significant, the node in $\mathbf{H}$ is considered important in the global context information learning and critical for distinguishing the semantic changes.

Local importance evaluation. The core idea is to identify the critical nodes $\boldsymbol{x}_i$ in $\mathbf{H}$ that are either representative of local emotions or indicative of significant local emotion shifts. Here we use the aforementioned window size to define the sub-graph $\mathbf{S}$ as the range of local context. The representation of the sub-graph can be formulated as:

$$\mathbf{h_S} = \text{Aggregation}\left(\boldsymbol{x}_i \mid i \in \mathcal{N}_w(v) \cap V\right), \quad (9)$$

where $\mathcal{N}_w(v)$ represents the neighborhood of node v defined by the window size w. Similar to global importance evaluation, we can obtain mutual information $I_L(\mathbf{h}^{'}_{\mathbf{S}}, \mathbf{h_S})$.

The overall importance of a node can be defined as:

$$I(v) = \gamma \tilde{I_G} + (1 - \gamma)\tilde{I_L} \quad (10)$$

where $\tilde{}$ is the normalization. $\gamma \in [0, 1]$ is the weight coefficient. We set lower weights for edges of $\rho\%$ nodes with the

least overall importance to create the augmented graph view $\mathbf{H}^{'}$. In this way, we introduce differences between the original and augmented views and preserve the graph's global and local characteristics, facilitating GCL.

Classifier

We use cross-entropy loss for classification as:

$$\mathcal{L}_{ce} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_i^j \log\left(\hat{y}_i^j\right), \tag{11}$$

where k is the number of emotion classes, n is the number of utterances, $\hat{y}_i^j$ is the i-th predicted label, and y_i^j is the i-th ground truth of j-th class. Our final objective function is:

$$\mathcal{L}_{all} = \mathcal{L}_{\text{ce}} + \lambda\mathcal{L}_{\text{cl}}, \tag{12}$$

λ is the weight of GCL loss and tuned on the validation set.

	Conversations			Utterances			
Dataset	**train**	**val**	**test**	**train**	**val**	**test**	**classes**
MELD	1,039	114	280	9,989	1,109	2,610	7
IEMOCAP	120	–	31	5,810	–	1,623	6

Table 1: Statistics of MELD and IEMOCAP datasets.

Experiments

Experimental Settings

Datasets and Metrics. We do evaluations on two renowned conversational datasets: 1) The IEMOCAP dataset (Busso et al. 2008) comprises scripted conversations performed by actors. Each utterance is classified into one of the following categories: anger, excitement, sadness, happiness, frustration, or neutral. As IEMOCAP lacks a predefined validation split, we adopt the approach of (Zhang, Chen, and Chen 2023; Li et al. 2023b), reserving the final 10% of the training set for validation. 2) The MELD dataset (Poria et al. 2019) is derived from multi-party conversations in the Friends TV series. Utterances are labeled with one of seven emotions: anger, disgust, sadness, joy, surprise, fear, or neutral. The statistics of these datasets are drawn in Table 1.

Implementation and hyperparameter setting. We employ the Weighted F1-score (WF1) as the evaluation metrics following Li et al. (2023b); Tu et al. (2024). All results are the average of 5 runs with different random seeds. We train models on a single RTX 3090 GPU based on the PyTorch framework. We tuned the hyperparameters for BIG-FUSION on the IEMOCAP and MELD datasets using their validation sets. We train BIG-FUSION using the Adam optimizer with a learning rate of 3e-5, a weight decay of 1e-8, and a batch size of 32 conversations. We create two different views for GCL with $\rho\%$ of 0.4 and 0.2 for the IEMOCAP dataset, and 0.6 and 0.3 for the MELD dataset, respectively. The temperature of GCL is 1, the sliding-window is 5, and the coefficient of GCL loss λ is 0.5. We set the size of the utterance window to 6 for the IEMOCAP dataset and 4 for the MELD dataset.

Baselines. For a comprehensive performance evaluation, we compare BIG-FUSION with the following SOTA baselines, categorized into **sequence-based** methods: DialogueRNN (Majumder et al. 2018), HCL (Yang et al. 2022), MVN (Ma et al. 2022), and SCMM (Yang et al. 2023); **graph-based fusion-based** methods: MMGCN (Hu et al. 2021), MMDFN (Hu et al. 2022), GA2MIF (Li et al. 2022), and CMCF-SRNet (Zhang and Li 2023); **sequential-design-based** methods: COGMEN (Joshi et al. 2022), Joyful (Li et al. 2023b), and AdaIGN (Tu et al. 2024).

Quantitative Results

With the same data partition, we evaluate the proposed BIG-FUSION against SOTA baselines to illustrate the effectiveness of our method. We report weighted F1 scores as the overall evaluation metric for MERC models and also provide F1 scores for each emotion class. Following Li et al. (2023b); Tu et al. (2024) we exclude the Fear and Disgust classes on the MELD dataset due to insufficient training samples to yield statistically significant results. The results of both two datasets are presented in Table 2. Specifically, we have the following findings: 1) BIG-FUSION demonstrates superior performance in the MERC task compared to SOTA sequence-, graph-, sequential design-based baselines. It achieves weighted F1 scores that surpass the strongest baseline by 2.65% and 0.57% on the two datasets, separately. 2) BIG-FUSION demonstrates improvements over the sequential-design-based baselines on these two datasets, which is attributed to its brain-inspired global-local context fusion framework. Unlike sequential designs that introduce bias by allowing global context to influence local context extraction, BIG-FUSION employs a dual-attention Transformer. This mechanism simultaneously processes global and local context information, preserving the independence and diversity of both contexts. By incorporating spiking dynamics into the attention module, BIG-FUSION mimics human neural processing, enhancing multimodal integration. The dual-evaluation approach in graph augmentation considers both overarching patterns and local structures, creating augmented graph views with comprehensive semantic representations. By addressing the inherent limitations of sequential designs, BIG-FUSION achieves a holistic understanding of emotion context in conversations, benefiting the MERC task.

Graph Contrastive Learning Methods

Here, we analyze the superiority of BIG-FUSION compared to other advanced GCL methods: GraphCL (You et al. 2020), GCA (Zhu et al. 2021), Auto-GCL (Yin et al. 2021), GCS (Wei et al. 2023), VMA (Duan et al. 2023) and AdGCL (Suresh et al. 2024). To ensure a fair comparison, we only replace the GCL component of BIG-FUSION and report the performance using optimal hyperparameter settings, and Table 3 shows results. This improvement is attributed to its dual-evaluation approach for graph augmentation in GCL, which accounts for node importance in both local and global contexts. By preserving nodes with crucial semantic significance in either context, BIG-FUSION creates augmented graph views that retain both global and local context

Methods	IEMOCAP							MELD					
	Happy	Sad	Neutral	Angry	Excited	Frustrated	WF1	Neutral	Surprise	Sadness	Happy	Anger	WF1
DialogueRNN	32.20	80.26	57.89	62.82	73.87	59.76	62.89	76.97	47.69	20.41	50.92	45.52	57.66
HCL	48.97	82.21	68.08	66.72	69.43	**68.73**	68.73	-	-	-	-	-	63.89
MVN	55.75	73.30	61.88	65.96	69.50	64.21	65.44	76.65	53.18	21.82	53.62	42.55	59.03
SCMM	45.37	78.76	63.54	66.05	76.70	66.18	67.53	-	-	-	-	-	59.44
MMGCN	45.14	77.16	64.36	68.82	74.71	61.40	66.26	76.33	48.15	26.74	53.02	46.09	58.31
MMDFN	42.22	78.98	66.42	69.77	75.56	66.33	68.18	77.76	50.69	22.93	54.78	47.82	59.46
GA2MIF	46.15	**84.50**	68.38	70.29	75.99	66.49	70.00	76.92	49.08	27.18	51.87	48.52	58.94
CMCF-SRNet	52.20	80.90	68.80	**70.30**	76.70	61.60	69.60	77.20	52.90	36.00	55.80	43.90	61.80
COGMEN	51.91	81.72	68.61	66.02	75.31	58.23	67.63	75.31	46.75	33.52	54.98	45.81	58.66
Joyful	60.94	84.42	68.24	69.95	73.54	67.55	71.03	76.80	51.91	41.78	56.89	50.71	61.77
AdaIGN	53.04	81.47	71.26	65.87	76.34	67.79	70.74	79.75	60.53	**43.70**	64.54	**56.15**	66.79
BIG-FUSION	**62.03**	83.64	**73.03**	68.38	**77.21**	68.66	**72.91**	**80.61**	**60.59**	41.82	**64.65**	55.59	**67.17**

Table 2: Performance comparison with baselines on the IEMOCAP and MELD datasets. The second-highest value is underlined.

Methods	IEMOCAP							MELD					
	Happy	Sad	Neutral	Angry	Excited	Frustrated	WF1	Neutral	Surprise	Sadness	Happy	Anger	WF1
GraphCL	54.49	80.25	70.32	69.79	70.91	65.03	69.01	79.63	57.10	40.63	65.23	50.54	65.09
GCA	53.16	82.17	73.22	66.30	72.73	67.69	70.68	80.02	57.52	42.04	64.44	53.64	66.31
GCS	51.59	80.81	72.09	68.62	70.99	66.03	69.60	79.34	58.68	38.55	64.55	55.60	66.20
AD-GCL	50.65	82.05	71.26	65.70	74.66	68.55	70.47	79.75	57.33	**42.14**	63.63	50.59	65.48
Auto-GCL	53.97	81.06	72.46	**69.59**	76.27	68.09	71.49	80.08	58.50	37.46	64.55	54.05	66.17
VMA	51.30	82.91	71.92	65.36	71.51	65.75	69.54	79.64	58.00	41.57	63.71	53.92	66.24
BIG-FUSION	**62.03**	**83.64**	**73.03**	68.38	**77.21**	**68.66**	**72.91**	**80.61**	**60.59**	41.82	**64.65**	**55.59**	**67.17**

Table 3: Performance comparison of different automated graph augmentation methods on IEMOCAP and MELD datasets.

Methods	**IEMOCAP**		**MELD**	
	Acc	WF1	Acc	WF1
COMPLETE	**72.64**	**72.91**	**68.24**	**67.17**
w/o GCL	68.39	68.50	65.51	64.79
- w/o $Local$	70.92	71.23	66.05	65.49
- w/o $Global$	71.60	71.76	66.93	65.99
w/o Spiking attention	71.53	71.72	67.13	65.93
- w/o $Local$	71.66	71.90	66.59	66.29
- w/o $Global$	69.44	69.82	66.74	65.82
w/o Threshold reg.	71.87	72.06	67.44	66.34

Table 4: Component-wise ablation study.

information, enabling GCL to learn comprehensive representations for the downstream emotion recognition task.

Ablation Study

To demonstrate the effectiveness of each component in BIG-FUSION, we conduct an ablation study using the IEMOCAP and MELD datasets. Following Tu et al. (2024), we report weighted F1 scores and Accuracy, with results presented in Table 4. The ablation study reveals that removing any component of BIG-FUSION leads to performance degradation, underscoring each part's significance in the MERC task. Notably, discarding either global or local context information in the Transformer or GCL components introduces performance drop, highlighting the necessity of considering both perspectives. Interestingly, local and global context information exhibit varying contributions in the Transformer and GCL components. This difference may reflect the inherent characteristics of Transformer and GCL in capturing distinct aspects of conversational context. We further investigate this phenomenon in the following section.

Global-Local Context Sensitive Study

In this section, we explore the importance of global and local context information in both Transformer and GCL components with IEMOCAP and MELD datasets, i.e. the sensitive study of α. The results are visualized in Figure 4. In both datasets, the weighted F1 scores show a general upward trend as the Local Weight for GCL and Global Weight for the Transformer increase. Peak performance is observed when both weights approach 0.6-0.8, indicating that a slight bias towards GCL's local information processing and Transformer's global information processing yields optimal results, aligning with the inherent strengths of these components. Notably, completely neglecting either aspect (weight of 0) leads to performance degradation, underscoring the im-

portance of considering both local and global context information. Overall, the optimal performance point occurs when each component focuses primarily on its specialization (local for GCL, global for Transformer) while maintaining consideration of another aspect of information.

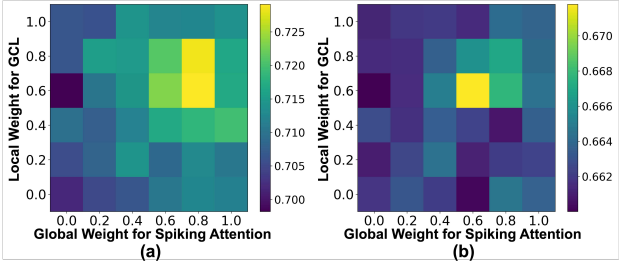

Figure 3: Heat maps illustrating the impact of balancing local and global information in Spiking Attention and GCL components for (a) IEMOCAP and (b) MELD datasets.

Case Study

To further demonstrate the effectiveness of BIG-FUSION, we visualize a case from the MELD dataset in Figure 5 and select Joyful and AdaIGN for comparison. Their incorrect prediction might stem from a sequential global-local context extraction mechanism. It may lead to an overemphasis on global information (overall emotion is sadness), causing it to interfere with and potentially override the neutral tone of the local context (the conversation turns), resulting in misclassification. In contrast, BIG-FUSION is brain-inspired and simultaneously extracts and preserves both global and local context. This parallel processing allows the model to maintain a balance between the overarching topic and the specific utterance content. By doing so, it can accurately capture the neutral tone of the local context without being overly influenced by the global sadness context, demonstrating its superior ability to integrate diverse context information.

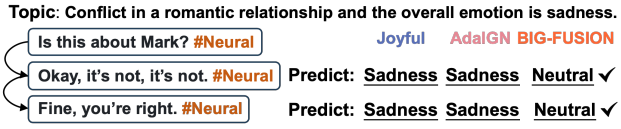

Figure 4: Emotion predictions across different models.

Effectiveness of Spiking Attention

In this section, we further explore the effectiveness of spiking attention from two aspects. Principal component analysis: as Figure 6 (a) shows, it achieves higher cumulative explained variance across principal components, particularly initial ones, indicating more effective concentration of key information in fewer dimensions. Figure 6 (b)'s steeper early curve for spiking attention confirms its enhanced ability to capture data variance with fewer components. This highlights its information extraction ability and the characteristic of condensing key information into fewer dimensions, which might be viewed as the interaction between different modalities; Optimization: From Figure 6 (c), the spiking contours are more concentrated and uniform near the local minimum, suggesting a more stable convergence to this point during the optimization process. Figure 6 (d) demonstrates that spiking achieves a lower loss and converges to a deeper local minimum, likely due to its superior feature extraction capability.

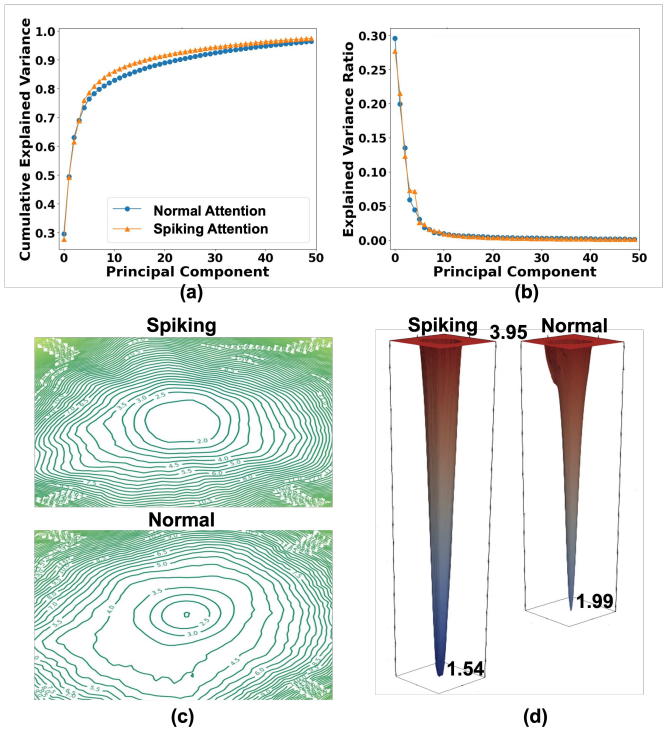

Figure 5: Principal component analysis (visualizations (a) and (b), alongside optimization landscapes (c) and (d), provide a comprehensive comparison between standard attention mechanisms and those based on spiking dynamics. These analyses are conducted using the IEMOCAP dataset, highlighting differences in feature extraction and optimization properties between the two approaches.

Conclusion

In this work, we propose BIG-FUSION, a brain-inspired global-local context fusion framework designed to overcome the challenges of current GCL-based sequential designs in the MERC task. BIG-FUSION incorporates a dual-attention Transformer and a dual-evaluation method for graph augmentation in GCL, enabling the simultaneous processing of global and local context information. This integration effectively balances context information processing and enhances the learning of comprehensive emotion semantics. By incorporating spiking dynamics, BIG-FUSION not only mimics brain neural functions but also improves the integration of multimodal information, enhancing feature extraction and increasing biological plausibility. Extensive experiments on two MERC datasets show that BIG-FUSION surpasses existing methods, proving its effectiveness.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No.62204140, Grant No. 62476035 and 62206037; in part by CAS Project for Young Scientists in Basic Research (YSBR-116); in part by National Natural Science Foundation of China (62325603, 62236009); in part by Beijing Science and Technology Plan (Z241100004224011).

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, E. A.; Provost, E. M.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359.

Duan, H.; Xie, C.; Li, B.; and Tang, P. 2023. Self-supervised contrastive graph representation with node and graph augmentation. *Neural Networks*, 167: 223–232.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 1459–1462. New York, NY, USA: Association for Computing Machinery. ISBN 9781605589336.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Conference on Empirical Methods in Natural Language Processing*.

Hu, D.; Hou, X.; Wei, L.; Jiang, L.; and Mo, Y. 2022. 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)*, 7037–7041.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the ACL-IJCNLP 2021 (Volume 1: Long Papers)*, 5666–5675. Online: Association for Computational Linguistics.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2016. Densely Connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

Jiang, T.; Xu, Q.; Ran, X.; Shen, J.; Lv, P.; Zhang, Q.; and Pan, G. 2024. Adaptive deep spiking neural network with global-local learning via balanced excitatory and inhibitory mechanism. In *The Twelfth International Conference on Learning Representations*.

Joshi, A.; Bhat, A.; Jain, A.; Singh, A.; and Modi, A. 2022. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4148–4164. Seattle, United States: Association for Computational Linguistics.

Li, D.; Tan, S.; Wang, Y.; Funakoshi, K.; and Okumura, M. 2023a. Temporal and Topological Augmentation-based Cross-view Contrastive Learning Model for Temporal Link Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 4059–4063. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701245.

Li, D.; Wang, Y.; Funakoshi, K.; and Okumura, M. 2023b. Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimoda Emotion Recognition. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16051–16069. Singapore: Association for Computational Linguistics.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2022. GA2MIF: Graph and Attention Based Two-Stage Multi-Source Information Fusion for Conversational Emotion Detection. *IEEE Transactions on Affective Computing*, 15: 130–143.

Lin, X.; Ma, L.; Liu, W.; and Chang, S.-F. 2019. Context-Gated Convolution. In *European Conference on Computer Vision*.

Liu, R.; Hu, Y.; Ren, Y.; Yin, X.; and Li, H. 2024. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18698–18706.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

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

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2018. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *AAAI Conference on Artificial Intelligence*.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics.

Primativo, S.; and Arduino, L. S. 2023. Global and Local Processing of Letters and Faces: The Role of Visual Focal Attention. *Brain Sciences*, 13(3).

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1): 61–80.

Schlichtkrull, M.; Kipf, T.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2017. Modeling Relational Data with Graph Convolutional Networks. In *Extended Semantic Web Conference*.

Sun, Y.; Yu, N.; and Fu, G. 2021. A Discourse-Aware Graph Neural Network for Emotion Recognition in Multi-Party Conversation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the EMNLP 2021*, 2949–2958. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Suresh, S.; Li, P.; Hao, C.; and Neville, J. 2024. Adversarial graph augmentation to improve graph contrastive learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713845393.

Tan, S.; Li, D.; Jiang, R.; Zhang, Y.; and Okumura, M. 2024. Community-Invariant Graph Contrastive Learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Tu, G.; Xie, T.; Liang, B.; Wang, H.; and Xu, R. 2024. Adaptive Graph Learning for Multimodal Conversational Emotion Detection. In *AAAI Conference on Artificial Intelligence*.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Neural Information Processing Systems*.

Wang, Q.; Fan, C.; Jia, T.; Han, Y.; and Wu, X. 2024. NDMRM: Neuronal Diversity Inspired Multisensory Recognition Model. In *AAAI Conference on Artificial Intelligence*.

Wang, Y.; Wang, X.; Qu, H.; Zhang, Y.; Chen, Y.; and Luo, X. 2021. Bio-inspired Model Based on Global-Local Hybrid Learning in Spiking Neural Network. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Wei, C.; Wang, Y.; Bai, B.; Ni, K.; Brady, D.; and Fang, L. 2023. Boosting Graph Contrastive Learning via Graph Contrastive Saliency. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th ICML*, volume 202 of *Proceedings of Machine Learning Research*, 36839–36855. PMLR.

Wu, Y.; Zhao, R.; Zhu, J.; Chen, F.; Xu, M.; Li, G.; Song, S.; Deng, L.; Wang, G.; Zheng, H.; Ma, S.; Pei, J.; Zhang, Y.; Zhao, M.; and Shi, L. 2020. Brain-inspired global-local learning incorporated with neuromorphic computing. *Nature Communications*, 13.

Wysoski, S. G.; Benusková, L.; and Kasabov, N. K. 2010. Brain-Like Evolving Spiking Neural Networks for Multimodal Information Processing. In *Brain-Inspired Information Technology*.

Xiao, T.; Wang, X.; Efros, A. A.; and Darrell, T. 2020. What Should Not Be Contrastive in Contrastive Learning. *International Conference on Learning Representations*, abs/2008.05659.

Xu, M.; Wu, Y.; Deng, L.; Liu, F.; Li, G.; and Pei, J. 2021a. Exploiting spiking dynamics with spatial-temporal feature normalization in graph learning. *arXiv preprint arXiv:2107.06865*.

Xu, Q.; Li, Y.; Fang, X.; Shen, J.; Liu, J. K.; Tang, H.; and Pan, G. 2023. Biologically inspired structure learning with reverse knowledge distillation for spiking neural networks. *arXiv preprint arXiv:2304.09500*.

Xu, Q.; Shen, J.; Ran, X.; Tang, H.; Pan, G.; and Liu, J. K. 2021b. Robust transcoding sensory information with neural spikes. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5): 1935–1946.

Yang, H.; Gao, X.; Wu, J.; Gan, T.; Ding, N.; Jiang, F.; and Nie, L. 2023. Self-adaptive Context and Modal-interaction Modeling For Multimodal Emotion Recognition. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the ACL 2023*, 6267–6281. Toronto, Canada: Association for Computational Linguistics.

Yang, L.; Shen, Y.; Mao, Y.; and Cai, L. 2022. Hybrid Curriculum Learning for Emotion Recognition in Conversation. *AAAI Conference on Artificial Intelligence*, abs/2112.11718.

Yin, H.; Zheng, H.; Mao, J.; Ding, S.; Liu, X.; Xu, M.; Hu, Y.; Pei, J.; and Deng, L. 2024. Understanding the functional roles of modelling components in spiking neural networks. *Neuromorphic Computing and Engineering*.

Yin, Y.; Wang, Q.; Huang, S.; Xiong, H.; and Zhang, X. 2021. AutoGCL: Automated Graph Contrastive Learning via Learnable View Generators. In *AAAI Conference on Artificial Intelligence*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph Contrastive Learning with Augmentations. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 5812–5823. Curran Associates, Inc.

Yu, F.; Wu, Y.; Ma, S.; Xu, M.; Li, H.; Qu, H.; Song, C.; Wang, T.; Zhao, R.; and Shi, L. 2023. Brain-inspired multimodal hybrid neural network for robot place recognition. *Science Robotics*, 8(78): eabm6996.

Zhang, D.; Chen, F.; and Chen, X. 2023. DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7395–7408. Toronto, Canada: Association for Computational Linguistics.

Zhang, X.; and Li, Y. 2023. A Cross-Modality Context Fusion and Semantic Refinement Network for Emotion Recognition in Conversation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13099–13110. Toronto, Canada: Association for Computational Linguistics.

Zhang, X.; Sun, J.; Hong, S.; and Li, T. 2024. Amanda: Adaptively Modality-Balanced Domain Adaptation for Multimodal Emotion Recognition. In *Findings of the ACL 2024*, 14448–14458.

Zhao, R.; Yang, Z.; Zheng, H.; Wu, Y.; Liu, F.; Wu, Z.; Li, L.; Chen, F.; Song, S.; Zhu, J.; et al. 2022. A framework for the general design and computation of hybrid neural networks. *Nature communications*, 13(1): 3427.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2021. Graph Contrastive Learning with Adaptive Augmentation. In *Proceedings of the Web Conference 2021*, WWW '21, 2069–2080. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.