Tracing Intricate Cues in Dialogue: Joint Graph Structure and Sentiment Dynamics for Multimodal Emotion Recognition

Jiang Li⬥ , Xiaoping Wang⬥ , *Senior Member, IEEE*, and Zhigang Zeng⬥ , *Fellow, IEEE*

Abstract—Multimodal emotion recognition in conversation (MERC) has garnered substantial research attention recently. Existing MERC methods face several challenges: (1) they fail to fully harness direct inter-modal cues, possibly leading to less-than-thorough cross-modal modeling; (2) they concurrently extract information from the same and different modalities at each network layer, potentially triggering conflicts from the fusion of multi-source data; (3) they lack the agility required to detect dynamic sentimental changes, perhaps resulting in inaccurate classification of utterances with abrupt sentiment shifts. To address these issues, a novel approach named GraphSmile is proposed for tracking intricate emotional cues in multimodal dialogues. GraphSmile comprises two key components, i.e., GSF and SDP modules. GSF ingeniously leverages graph structures to alternately assimilate inter-modal and intra-modal emotional dependencies layer by layer, adequately capturing cross-modal cues while effectively circumventing fusion conflicts. SDP is an auxiliary task to explicitly delineate the sentiment dynamics between utterances, promoting the model's ability to distinguish sentimental discrepancies. GraphSmile is effortlessly applied to multimodal sentiment analysis in conversation (MSAC), thus enabling simultaneous execution of MERC and MSAC tasks. Empirical results on multiple benchmarks demonstrate that GraphSmile can handle complex emotional and sentimental patterns, significantly outperforming baseline models.

Index Terms—Emotion recognition in conversation, sentiment dynamics, multimodal fusion, graph neural networks.

I. INTRODUCTION

EMOTION recognition in conversation (ERC) has become a hot topic in academia due to its broad application prospects in various fields such as government affairs, education, and healthcare. A multitude of ERC models [1]–[4] focus on the analysis of textual modality, often encountering the issue of insufficient information. In certain special cases, such as sarcasm, text-based methods may not fully comprehend the semantics of an utterance from solely a single modality. The extraordinary ability of the human brain to integrate multisensory information provides inspiration for addressing the limitations of information deficiency in unimodal models. Multimodal emotion recognition in conversation (MERC) emerges as an effective solution to this challenge, achieving more accurate emotional state recognition by integrating information from various modalities such as text, audio, and video. MERC is a sophisticated task that goes beyond the scope of traditional unimodal [5], [6] and non-conversational [7], [8] affective endeavors (in this paper, the term "affective" encompasses emotional and sentimental aspects). It necessitates not only the sufficient extraction of cross-modal association cues that capture synergies between different sensory inputs (e.g., facial expressions, vocal tones, and textual content), but also an in-depth investigation of intra-modal contextual cues to perceive effects of contexts on the emotional state of an utterance. Compared to unimodal and non-conversational counterparts, MERC faces more formidable challenges, with the key one being the thorough capture of intra-modal contextual cues and inter-modal associative cues.

Researchers have developed numerous MERC approaches, including new architectures and learning mechanisms, to capture and synthesize the wealth of information present in multimodal dialogues. Some efforts [9]–[12] straightforwardly concatenated or summed features across disparate modalities, which have manifested certain performance merits over unimodal approaches when dealing with specific tasks. Regrettably, certain modalities may convey information that is ambiguous or even contradictory to labels. These concatenation fusion-based works fail to adequately account for the intrinsic relationships and interactions between different modalities, thus making it difficult to effectively address the issues of information selection and integration. There are an array of methods [13]–[16] predicated on cross-attention fusion. These methods achieve the amalgamation of diverse modality information by calculating the weights between multimodal features and applying these weights to the current modality

features. Nonetheless, they retain certain constraints at the feature level and have not fully actualized direct interaction and fusion between different modalities.

Recent studies have unveiled the exceptional potential of graph neural networks (GNNs) in handling the fusion and comprehension of multimodal data. GNNs, with the unique topological structure and dynamic information propagation mechanism, inherently excel at capturing cross-modal associative cues and intra-modal contextual cues. Despite significant advancements [17]–[20] in the graph-based MERC domain, several unresolved issues persist. Firstly, existing models, when constructing cross-modal graphs, only connect heterogeneous nodes (i.e., nodes that belong to different modalities) from the same utterance and fail to incorporate associations with heterogeneous nodes from different utterances. This construction scheme overlooks the direct emotional cues provided by other utterances across modalities, leading to incomplete capture of multimodal information. Secondly, these models typically process information from the same and different modalities at each layer of the network, i.e., concurrently extract intra-modal and inter-modal emotional cues. Owing to the heterogeneity among modalities [18], [21], [22], this concurrent aggregation of different types of information may precipitate conflicts during fusion, resulting in insufficient cross-modal modeling. Furthermore, current models are excessively preoccupied with contextual modeling, neglecting the dynamic mutation of sentiments throughout the conversation. The absence of agility in recognizing sentimental changes may cause these models' inability to classify accurately in scenarios with abrupt sentiment shifts. Consequently, there is an urgent need to develop models that can integrate multimodal heterogeneous information more comprehensively and rationally, and that are capable of keenly capturing the sentiment dynamics.

To cope with the above dilemmas, we introduce a joint graph structure and sentiment dynamics for multimodal emotion recognition in conversation (GraphSmile). The core components of GraphSmile encompass the graph structure fusion (GSF) and sentiment dynamics perception (SDP) modules. During the construction of cross-modal edges, GraphSmile not only correlates the current node with heterogeneous nodes belonging to the same utterance but also directly connects it with nodes emanating from disparate utterances. This strategy for edge construction effectively ameliorates the insufficient in extracting cross-modal information. At the graph convolution stage, the GSF module adeptly utilizes the graph structure to alternately collect inter-modal associative cues and intra-modal contextual cues on layer-by-layer basis, thereby precluding potential conflicts stemming from the simultaneous aggregation of diverse data sources at each layer. To mitigate redundant connections in the multimodal dialogue graph, we refrain from constructing edges between nodes within the same modality; nonetheless, the GSF module, through its unique graph convolution mechanism, is capable of directly and comprehensively capturing intra-modal contextual cues. In the phase of sentiment dynamics detection, we integrate the SDP module as an auxiliary task to explicitly model the shift in sentimental state between utterances. This module not only deepens the model's comprehension of sentiment

dynamics but also constrains the network parameters of the main task with a strategy akin to contrastive learning, i.e., drawing utterances with the same sentiment closer and those with different sentiments further apart, thereby enhancing the model's ability to discern sentimental disparities.

In this work, we can effortlessly extend GraphSmile into a multitask model. Namely, GraphSmile is applied to the multimodal sentiment analysis in conversation (MSAC) task and achieves multitask joint optimization with the MERC task. Specifically, we initially incorporate a sentiment classifier, followed by the joint formulation of emotion classification loss, sentiment classification loss, and shift classification loss to construct a multitask learning objective, thereby synergistically optimizing network parameters. In this manner, our GraphSmile is capable of efficiently executing both MSAC and MERC tasks. In the domain of affective computing, emotions (e.g., happiness, surprise, neutrality, sadness, and anger) are typically regarded as more nuanced and fleeting psychological states, characterized by their situational and temporary nature, while sentiments (e.g., negative, neutral, positive) are understood as coarser and more enduring psychological inclinations, marked by their profundity and stability. While emotion and sentiment exhibit varying levels of granularity, duration, and characteristic, there exists an interrelation between them; happiness, for example, is typified as a positive sentiment, whereas sadness is reflective of a negative sentimental disposition. Thus, our co-optimization strategy not only facilitates knowledge transfer and complementary advantages between different granularity levels of affective tasks but also significantly enhances GraphSmile's comprehensive perceptive capability for emotional and sentimental cues.

The principal contributions of this work are as follows:

- A multitask multimodal affective model that unites graph structure and sentiment dynamics, GraphSmile, is proposed for both MERC and MSAC tasks. GraphSmile exploits multiple losses to achieve multitask synergistic optimization while upgrading its ability to synthesize the perception of various cues.
- The GSF module we designed not only ensures the full exploitation of cross-modal affective cues but also obviates potential conflicts arising from the fusion of multi-source data. This module also confers two additional benefits, i.e., it diminishes redundant connections within the graph and alleviates the over-smoothing issue prevalent in graph-based models.
- We develop a plug-and-play SDP module that can explicitly capture sentimental changes in dialogues, fostering the model's understanding of the dynamics of sentimental evolution while constraining the network parameters of the primary task.
- Extensive comparative experiments and ablation studies have been conducted on multimodal benchmark datasets. The empirical results have validated the superiority of GraphSmile in both MERC and MSAC tasks.

The remainder of this paper is organized as follows. Section II reviews the related efforts. Section III provides a detailed exposition of the proposed GraphSmile. Sections IV

and V constitute the experimental segment of our work. Section VI summarizes this work.

II. RELATED WORK

A. Emotion Recognition in Conversation

With the proliferation of applications for dialogue systems, ERC has increasingly become a focal point in the realm of affective computing. Depending on the source of input information, research in this domain can be categorized into two branches: unimodal and multimodal approaches. Since the inception of ERC tasks, text-based methods [23]–[27] have consistently held a dominant position among unimodal approaches. In recent years, this branch has seen profound advancements with numerous emerging technologies modeling contextual information in textual conversations from diverse perspectives. GNNs, due to their proficiency in capturing intra-modal dependency information, have notably been instrumental in conversational relationship modeling, thereby giving rise to numerous graph-based ERC models [28]–[33]. Regrettably, unimodal ERC methods are limited by their inherent information deficiencies and susceptibility to external influences, which can impede their performance.

The richness of human emotions manifests distinctively across individual modalities such as text, audio, and vision. These modalities possess unique characteristic and exhibit both shared traits and individual nuances. Consequently, multimodal ERC techniques have garnered significant attention due to their capacity to integrate diverse modal information. Some methods [9]–[12], [34] employ a kind of straightforward feature fusion strategy, which involves concatenating or summing feature vectors from different modalities, overlooking the intrinsic interactions between modalities. In tandem, a variety of models [14], [16], [35], [36] based on cross-modality attention are proposed in succession. These methods are unable to achieve direct absorption of information from one modality to another due to the constraints of their network structures, which are considered implicit cross-modal modeling.

In recent years, GNNs have garnered widespread attention due to their powerful capabilities in handling complex relationships, providing a solvable inspiration for explicitly conducting cross-modal interactions. Despite certain advancements in graph-based fusion methods [18], [19], [37], [38], they still face a series of unique and common issues: due to the limitations of graph construction strategies, they may not fully capture and utilize cross-modal emotional cues; the heterogeneity between different modal data may lead to conflicts in the fusion process of multi-source information; some methods may overly rely on extracting intra-modal contextual cues, neglecting the dynamic changes in sentimental states during the conversation. Moreover, some of the aforementioned issues also exist in methods based on concatenation fusion and cross-attention fusion. In response to these issues, we propose a new fusion network architecture called GraphSmile for MERC.

B. Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA), due to its wide range of application prospects, has become a focal research topic in the field of affective computing. MSA is a multimodal task aimed at understanding and detecting human emotional tendencies by processing multimodal heterogeneous data, such as text, speech, and vision. The main challenge of the MSA task is how to effectively harness information from different modalities to complement each other, and a plethora of representative works [39]–[42] have been proposed in succession. Existing mainstream works focus on enhancing the performance of MSA models by designing complex structures, interaction mechanisms, or fusion paradigms, which can be broadly categorized into tensor-based methods [39], [43], attention-based methods [41], [44], and graph-based methods [45], [46]. These research trajectories of MSA tasks are distinctly different from those of the ERC tasks, as they do not take into account the conversational context, rendering them incompatible with ERC tasks. Although Hu et al. [11] have made effort to unify ERC and MSA, their approach to handling labels in the datasets is overly complex and cannot fully guarantee the correctness of universal labels. Divergent from the extant MSA methods, our GraphSmile is conceived as a multitask multimodal conversational affective model. In GraphSmile, we only need to merge emotional labels to effortlessly unify the ERC and MSA tasks, thereby facilitating an efficacious excavation of both emotional and sentimental cues in a concurrent manner.

III. PROPOSED GRAPHSMILE

Given a conversation $c = [u_1, u_2, \ldots, u_M]$ consisting of M utterances, along with the emotional label e_i (or sentimental label s_i) corresponding to each utterance u_i, the objective of the MERC (or MSAC) task is to train a model based on the dialogue and labels to detect the affective state of each utterance. In MERC and MSAC tasks, each utterance comprises three modalities: textual t, visual v, and acoustic a modalities, which is denoted as $u_i = [u_i^t, u_i^v, u_i^a]$.

Effectively extracting complex affective cues from multimodal dialogues is crucial for the MERC task. Here, the affective cues mainly consist of inter-modal associative and intra-modal contextual cues. *Inter-Modal Associative Cues:* Assuming the existence of two modalities (e.g., textual and visual), an utterance u_i is represented differently across modalities as u_i^t and u_i^v. The information carried by u_i^t is considered as an inter-modal associative cue for u_i^v. The inter-modal associative cues for u_i^v also include information carried by other utterances from modality t, such as u_{i-2}^t, u_{i-1}^t, and u_{i+1}^t. *Intra-Modal Contextual Cues:* Taking a single modality (e.g., acoustic) as an example, an utterance u_j is denoted as u_j^a in a specific modality. The intra-modal contextual cues for u_j^a encompass information carried by other utterances from the same modality a, such as u_{j-1}^a, u_{j+1}^a, and u_{j+2}^a.

To address the predicaments encountered by existing models, we propose a model termed GraphSmile that conducts a comprehensive modeling of multimodal dialogues. The overall architecture of GraphSmile is depicted in Fig. 1, which mainly includes the extraction of utterance-level features, the construction of multimodal dialogue graphs, the alternate graph propagation, the multimodal integration, and the multitask
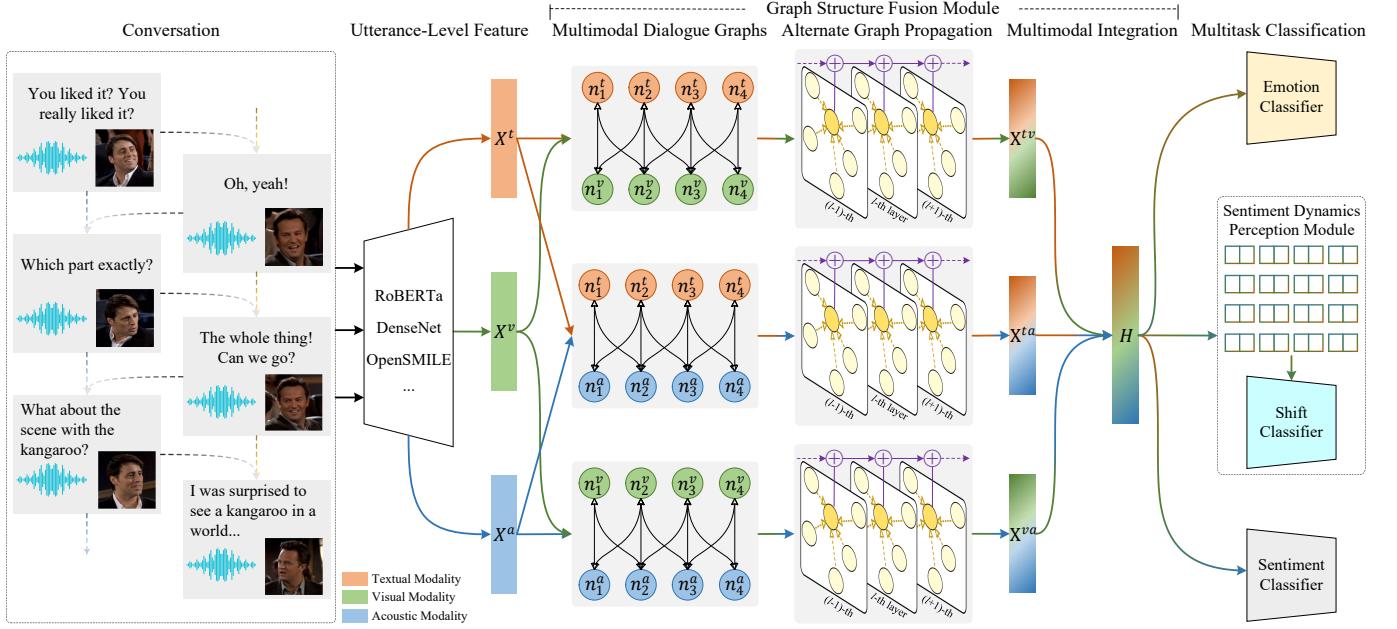
Fig. 1. Overall architecture of our GraphSmile. Initially, utterance-level features are extracted for each modality. Subsequently, three multimodal dialogue graphs are constructed. Then, alternating graph propagation is conducted to explore affective cues. Finally, multitask classification is performed.

classification. Here, multitask classification involves emotion, shift, and sentiment classifications. In the following, we describe each component of GraphSmile in detail.

A. Utterance-Level Feature

The datasets utilized in this work include IEMOCAP [47] (both 6-way and 4-way settings), MELD [48], and CMU-MOSEI [49]. Influenced by prior works [2], [17], [50], the utterance-level feature extraction for the datasets used in GraphSmile is as follows. Text features for all datasets are processed using RoBERTa [51]. Visual features for the IEMOCAP (6-way) and MELD datasets are extracted using DenseNet [52], which is pre-trained on the FER+ corpus [53]. The IEMOCAP (4-way) dataset utilizes OpenFace 2.0 [54] for visual feature extraction, while the visual features for the CMU-MOSEI dataset are derived from the work of Zadehet et al. [49]. Audio features for the IEMOCAP (6-way) and MELD datasets are extracted using the OpenSmile toolkit with IS10 configuration [55]. The IEMOCAP (4-way) dataset uses OpenSmile [56] for audio feature extraction, and the CMU-MOSEI dataset uses librosa [57] for this purpose.

B. Graph Structure Fusion Module

1) Multimodal Dialogue Graphs: To establish relationships between utterances while considering inter-modal interactions, we construct three multimodal dialogue graphs, each composed of two modalities. Taking the textual and visual modalities as an example, we build a textual-visual dialogue graph $\mathcal{G}^{tv} = (\mathcal{N}^t, \mathcal{N}^v, \mathcal{E}^{t \leftarrow v}, \mathcal{E}^{v \leftarrow t})$. Here, $\mathcal{N}^t$ and $\mathcal{N}^v$ represent the node sets from the textual and visual modalities, respectively. $\mathcal{E}^{t \leftarrow v}$ denotes the set of edges from source visual nodes to target textual nodes, while $\mathcal{E}^{v \leftarrow t}$ denotes the set of edges from source textual nodes to target visual nodes.

Nodes: Each utterance u_i is represented as nodes n_i^t for the textual modality and n_i^v for the visual modality, with corresponding features denoted as x_i^t and x_i^v. If a dialogue consists of m utterances, then $\mathcal{G}^{tv}$ contains $2m$ nodes. Consequently, $\mathcal{N}^t$ can be represented as $\{n_1^t, n_2^t, \ldots, n_M^t\}$, $\mathcal{N}^v$ can be represented as $\{n_1^v, n_2^v, \ldots, n_M^v\}$, with their corresponding features denoted as $[x_1^t, x_2^t, \ldots, x_M^t]$ and $[x_1^v, x_2^v, \ldots, x_M^v]$.

Edges: We establish relationships only between nodes of different modalities. In the graph $\mathcal{G}^{tv}$, a textual node n_i^t is connected to visual nodes; similarly, a visual node n_i^v is connected to textual nodes. To mitigate redundant connections [18], [58], we set up two sliding windows. Specifically, n_i^t is associated with the past P visual nodes (i.e., $n_{i-P}^v, n_{i-P+1}^v, \ldots, n_{i-1}^v$), and also with the future F visual nodes (i.e., $n_{i+1}^v, n_{i+2}^v, \ldots, n_{i+F}^v$). The same connection way is applied to the visual node n_i^v. Additionally, different modal nodes (e.g., n_i^t and n_i^v) from the same utterance are also connected.

Edge Weights: In our research, the edges connecting two nodes are directed, with edges of differing orientations signifying distinct implications. For instance, in the graph $\mathcal{G}^{tv}$, for nodes n_i^t and n_{i-1}^v, the edge $e(n_i^t \leftarrow n_{i-1}^v)$ denotes the informational contribution from n_{i-1}^v to n_i^t, whereas the edge $e(n_{i-1}^v \leftarrow n_i^t)$ signifies the informational contribution from n_i^t to n_{i-1}^v. To distinguish between different edges, diverging from existing methods [17], [18], we assign a set of trainable weights directly to them. These weights are initialized to a value of 1 and are optimized during the model training process.

Adjacency Matrix: The adjacency matrix serves to store the relationships between all nodes in a graph, representing a pivotal element in expressing the graph structure and also constituting one of the input data for spectral-based GNNs. Based on the nodes, edges, and edge weights previously established, an adjacency matrix A^{tv} can be generated to serve

as the input for our GraphSmile.

Following the construction manner of the graph $\mathcal{G}^{tv}$, a textual-acoustic dialogue graph $\mathcal{G}^{ta}$ and a visual-acoustic dialogue graph $\mathcal{G}^{va}$ can also be obtained. The majority of previous approaches [17], [18], [20], when creating cross-modal edges, only connect nodes from the same utterance, thereby neglecting the affective cues directly provided by other utterances. In contrast to these methods, in our construction of cross-modal edges, the current node is not only connected to nodes from the same utterance but also directly linked to nodes from different utterances. It is noteworthy that our multimodal dialogue graphs don't establish edges between utterances in the same modality. Nonetheless, GraphSmile can still effortlessly capture intra-modal contextual affective cues through the unique graph convolution mechanism. In Part III-B3, we explain how these intra-modal affective cues can be naturally extracted by graph convolution.

2) Alternate Graph Propagation: Previous efforts [17], [18], [20] have demonstrated that GNNs can effectively mine intra-modal contextual information and cross-modal associative information in multimodal dialogues. However, these models share a common issue: the simultaneous integration of inter-modal and intra-modal affective cues at each network layer may cause conflicts among multi-source information, making thorough cross-modal modeling difficult. This part will provide a detailed description of how the GSF module cleverly addresses this issue.

As depicted in Fig. 1, drawing inspiration from existing GNN works, we define a new graph convolutional operation to extract affective cues from multimodal dialogues. For each dialogue graph, multiple simplified graph convolutional layers [59] are utilized to propagate inter-modal and intra-modal cues layer by layer in an alternating fashion, which can be mathematically expressed as:

$$X^{(l)} = AX^{(l-1)}\Theta^{(l)}, \quad l = 1, 2, \ldots, L. \quad (1)$$

Here, A denotes the adjacency matrix of the dialogue graph, $X^{(l)}$ signifies the output at the l-th layer, and $X^{(0)}$ represents the initial input to the GSF module; $\Theta^{(l)}$ indicates the learnable parameter at the l-th layer. Assuming a given textual-visual dialogue graph $\mathcal{G}^{tv}$, then the adjacency matrix A can be represented by A^{tv}, $X^{(l)}$ can be denoted by $X^{tv,(l)}$. The initial feature $X^{tv,(0)}$ is composed of $X^{t,(0)}$ and $X^{v,(0)}$, where $X^{t,(0)}$ and $X^{v,(0)}$ are denoted as $[x_1^t, x_2^t, \ldots, x_M^t]$ and $[x_1^v, x_2^v, \ldots, x_M^v]$, respectively.

Subsequently, the feature representations obtained from all layers are summed together, which is known as residual connections, and can be mathematically formulated as:

$$X = X^{(0)} + \text{FC}^{(1)}(X^{(1)}) + \cdots + \text{FC}^{(L)}(X^{(L)}). \quad (2)$$

In the graph $\mathcal{G}^{tv}$, X can be represented as X^{tv}, which is composed of $X^{t \leftarrow v}$ and $X^{v \leftarrow t}$. Here, $X^{t \leftarrow v}$ denotes the textual feature representation that has incorporated visual information, and $X^{v \leftarrow t}$ signifies the visual feature representation that has carried textual information. FC($\cdot$) represents a fully connected layer, which is composed of a linear transformation, an activation function (e.g., LeakyReLU), and a dropout operation,

i.e.,

$$\text{FC}(X) = \text{Dropout}(\text{LeakyReLU}(X\Theta_f + b_f)). \quad (3)$$

Here, Θ_f and b_f are the learnable parameters. In our GSF module, the fully connected layer serves two purposes: (1) the trainable network parameters allow the GSF module to dynamically select information from each network layer; (2) the combination of the linear transformation and activation function can further enhance the expressive ability of the designed GSF module.

3) Example Explanation: We take the textual-visual dialogue graph $\mathcal{G}^{tv}$ as an example to elucidate why the GSF module can alternately propagate different types of affective cues layer by layer. In Fig. 2, we illustrate the process of multi-layer graph convolution with a concrete example involving four utterances.
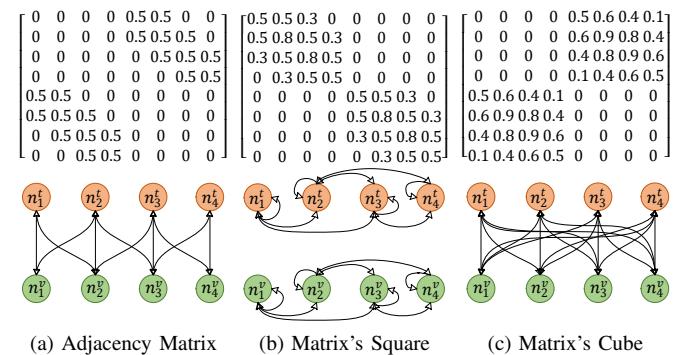
Fig. 2. Example explanation of alternating propagation in the GSF module. Here, (a), (b), and (c) represent the adjacency matrix A, the matrix's square A^2, and the matrix's cube A^3, along with their corresponding graph structures, respectively. In actual implementation, the edge weights in the dialogue graph $\mathcal{G}^{tv}$ are trainable parameters; for the sake of narrative simplicity, we assume all edge weights to be 0.5.

According to Eq. 1, the first-layer graph convolution process can be mathematically formulated as:

$$X^{(1)} = AX^{(0)}\Theta^{(1)}. \quad (4)$$

Here, A can be represented by A^{tv}, which is the adjacency matrix that expresses the relationships between all nodes in the graph $\mathcal{G}^{tv}$; $X^{(1)}$ denotes the output at the first layer, $X^{(0)}$ is the initial input to the GSF module, and $\Theta^{(1)}$ indicates the learnable parameter at the first layer. From the construction way of the multimodal dialogue graph (see Fig. 2a), A^{tv} does not include the relationships between nodes from the same modality, instead only storing the associations between nodes from different modalities. Consequently, during the first-layer graph convolution, each node only collects affective cues from different modal nodes. For instance, in Fig. 2a, the node n_1^t affiliated with the textual modality is exclusively connected to visual nodes (e.g., n_1^v, n_2^v), thereby assimilating cues from a distinct modality; reciprocally, the visual node n_1^v seizes affective cues solely from textual counterparts.

For the second-layer graph convolution, integrating Eq. 1 and Eq. 4, the message propagation process can be mathemat-

ically formalized as:

$$
\begin{aligned}
X^{(2)} &= AX^{(1)}\Theta^{(2)} \\
&= A(AX^{(0)}\Theta^{(1)})\Theta^{(2)} \\
&= A^2 X^{(0)}\Theta^{(1)}\Theta^{(2)}.
\end{aligned} \tag{5}
$$

Here, $X^{(2)}$ and $\Theta^{(2)}$ represent the output and learnable parameter for the second layer, respectively. Despite the second layer's features being predicated on those of the first layer, according to Eq. 5, the output of the second layer can also be directly computed from the initial input $X^{(0)}$. In our GSF module, since A^2 stores the second-hop neighborhood relationships of nodes, the significance of the second-layer graph convolution can be interpreted as nodes assimilating feature information from their second-hop neighbors. Referring to Fig. 2b, during the second-layer graph convolution, each node collects affective cues solely from nodes within the same modality. For instance, the textual node n_1^t is connected only to nodes belonging to the textual modality, thereby collecting cues exclusively from intra-modal nodes (e.g., n_2^t, n_3^t); similarly, the visual node n_1^v gathers affective cues only from visual nodes. Thus, the above message passing process explains why our GraphSmile can still naturally capture intra-modal affective cues without creating intra-modal edges.

The message propagation process for the third-layer graph convolution can be expressed as follows:

$$
\begin{aligned}
X^{(3)} &= AX^{(2)}\Theta^{(3)} \\
&= A(AX^{(1)}\Theta^{(2)})\Theta^{(3)} \\
&= A(A(AX^{(0)}\Theta^{(1)})\Theta^{(2)})\Theta^{(3)} \\
&= A^3 X^{(0)}\Theta^{(1)}\Theta^{(2)}\Theta^{(3)}.
\end{aligned} \tag{6}
$$

Here, $X^{(3)}$ signifies the feature matrix resulting from the third layer and $\Theta^{(3)}$ represents the corresponding learnable parameter. Following Eq. 6, the third-layer graph convolution, akin to the second layer, can have its output directly computed from the initial features $X^{(0)}$. A^3 contains the third-hop neighborhood relationships, thus the implication of the third-layer is that nodes integrate feature information from their third-hop neighbors. As depicted in Fig. 2c, during the third-layer graph convolution, each node absorbs affective cues solely from nodes of another modality. For instance, similar to the first-layer convolution, the textual node n_1^t is connected only to nodes of a different modality, i.e., visual nodes, and the visual node n_1^v is connected only to textual nodes.

Through the three examples mentioned above, it can be concluded that the first layer of graph convolution propagates affective cues across different modalities, the second layer of graph convolution conveys affective cues within the same modality, and the third layer of graph convolution once again captures inter-modal associative cues, and so on. Therefore, these examples illustrate that our GSF module can alternately propagate inter-modal and intra-modal affective cues layer by layer, thereby preventing conflicts among multi-source information. Summarizing the patterns from the examples, to capture affective cues from more distant neighbors, we can

extend to the l-th layer graph convolution:

$$
\begin{aligned}
X^{(l)} &= AX^{(l-1)}\Theta^{(l)} \\
&= A(AX^{(l-2)}\Theta^{(l-1)})\Theta^{(l)} \\
&= A\cdots(AX^{(0)}\Theta^{(1)})\cdots\Theta^{(l)} \\
&= (A\cdots A)X^{(0)}(\Theta^{(1)}\cdots\Theta^{(l)}) \\
&= A^l X^{(0)}(\Theta^{(1)}\cdots\Theta^{(l)}).
\end{aligned} \tag{7}
$$

Here, $X^{(l)}$ represents the feature matrix at the l-th layer, A^l denotes the l-th power of the adjacency matrix A, which signifies the reach to l-hop neighbors, and $\Theta^{(l)}$ are the learnable parameters specific to the l-th layer. This formulation allows nodes to aggregate information from increasingly broader neighborhoods as the layer index l increases, thus enriching the representation of each node with more comprehensive affective cues. Via Eq. 2, the GSF module is capable of integrating the outputs from all layers. The summation of all layer outputs, i.e., the residual connection, provides two ingenious benefits: (1) it mitigates the over-smoothing problem [60] in GNNs; (2) it amalgamates affective cues from multi-hop neighbors.

4) Multimodal Integration: The graphs $\mathcal{G}^{tv}$, $\mathcal{G}^{ta}$, and $\mathcal{G}^{va}$ are respectively input into three separate GSF modules for graph convolution operations. After stacking multiple layers of convolutions and combining with Eq. 2, three outputs X^{tv}, X^{ta}, and X^{va} can be obtained. Here, X^{tv} is constituted by $\mathrm{X}^{t\leftarrow v}$ and $\mathrm{X}^{v\leftarrow t}$, X^{ta} is composed of $\mathrm{X}^{t\leftarrow a}$ and $\mathrm{X}^{a\leftarrow t}$, X^{va} consists of $\mathrm{X}^{v\leftarrow a}$ and $\mathrm{X}^{a\leftarrow v}$. Then, these outputs are summed to obtain a new feature representation:

$$
\begin{aligned}
H = &\,\sigma((\mathrm{X}^{t\leftarrow v})\Theta_h) + \sigma((\mathrm{X}^{v\leftarrow t})\Theta_h) \\
&+ \sigma((\mathrm{X}^{t\leftarrow a})\Theta_h) + \sigma((\mathrm{X}^{a\leftarrow t})\Theta_h) \\
&+ \sigma((\mathrm{X}^{v\leftarrow a})\Theta_h) + \sigma((\mathrm{X}^{a\leftarrow v})\Theta_h),
\end{aligned} \tag{8}
$$

where σ represents the activation function, e.g., LeakyReLU, and Θ_h denotes the shared learnable parameter. The new representation H integrates information from textual-to-visual, textual-to-acoustic, and visual-to-acoustic modalities, providing comprehensive features that encapsulate the multimodal affective cues in dialogues.

C. Emotion Classification

To achieve emotion classification, H is fed into a classifier that incorporates a linear function followed by a softmax function. The emotion classification is formulated as:

$$
\begin{aligned}
\hat{Y} &= \texttt{softmax}(H\Theta_y + b_y), \\
\hat{e}_i &= \texttt{argmax}(\hat{y}_i[\tau]),
\end{aligned} \tag{9}
$$

where both Θ_y and b_y are learnable parameters. Let $\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M]$, then $\hat{y}_i$ represents the emotional probability vector of utterance u_i, and $\hat{e}_i$ denotes the corresponding predicted emotional label. During the training phase, we employ the categorical cross-entropy loss for network optimization, which can be mathematically expressed as:

$$
\mathcal{L}_e = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C_e} y_{i,j}\log \hat{y}_{i,j}, \tag{10}
$$

where M represents the number of utterances in the dialogue, C_e denotes the number of emotional categories; $y_{i,j}$ signifies the ground-truth probability distribution of the i-th utterance u_i in the j-th category, and $\hat{y}_{i,j}$ is the predicted probability of utterance u_i.

D. *Sentiment Dynamics Perception Module*

In dialogue systems, sentiment shift refers to the phenomenon where the sentimental state of two utterances transitions from one category to another. Capturing the dynamic change in sentiments is pivotal for comprehending the sentimental trajectory of a conversation. However, prevailing MERC models predominantly concentrate on contextual modeling of the utterance, exhibiting a dearth of acuity in detecting inter-utterance sentimental changes. This limitation can lead to inaccuracies in the classification of affective states during scenarios characterized by sentiment shifts. To mitigate this shortfall, we introduce an auxiliary task known as the sentiment dynamics perception (SDP) module. Our SDP module serves the following purposes: (1) it models the dynamics of sentimental changes, endowing the model with the capability to not only discern static affective states but also track dynamic sentimental shifts; (2) by pulling utterances with the same sentimental state closer and pushing those with different sentimental states further apart, it aids in enhancing the model's ability to discern between various affective categories; (3) integrated as a complementary task, its loss function can be conjointly optimized with that of the main task, guiding the model to accord significance to the phenomenon of sentiment shifts.

1) Shift Label: Based on the shifts of sentimental states between utterances in a dialogue, we generate a set of sentiment shift labels for that conversation. In a dialogue, if two utterances possess distinct sentiments, i.e., their states transition from one sentiment to another, then the shift label is set to "1", signifying that a sentiment shift has occurred between them. Conversely, if two utterances share the same sentiment, i.e., there is no alteration in their sentimental state, then the shift label is set to "0", indicating that no sentiment shift has transpired between them. The annotating process of the shift label $o_{(i,j)}$ can be formulated as follows:

$$o_{(i,j)} = \begin{cases} 1, & \text{if } s_i \neq s_j, \\ 0, & \text{if } s_i = s_j. \end{cases} \quad (11)$$

Here, s_i denotes the sentimental state of the i-th utterance u_i. It is worth noting that the IEMOCAP datasets do not provide pre-defined sentiment labels. To make them compatible with GraphSmile, we merge their emotion labels, with the specific merging scheme as depicted in Table I. Additionally, in the CMU-MOSEI dataset, an utterance may correspond to multiple emotion labels, and this annotation way is not suitable for our model. Consequently, we adopt seven sentiments, ranging from highly negative (-3) to highly positive (+3), to substitute the emotion labels for classification.

2) Shift Feature: In Subsection III-B4, a fused multimodal feature representation $H \in \mathbb{R}^{M \times D}$ is obtained. Here, M

denotes the number of utterances in the dialogue, D signifies the feature dimension, and H can be indicated as $[h_1, h_2, \ldots, h_M]$. Based on the feature representation H, we construct a set of sentiment shift features. Specifically, given a dialogue containing M utterances, we concatenate the features of these utterances pairwise. Through the aforementioned operation, M^2 samples of sentiment shifts can be generated to construct a shift feature representation T with a size of $M^2 \times 2D$.

Fig. 3 provides an example of generating shift features. Assuming there are four utterances in the dialogue, their features are represented as $[h_1, h_2, h_3, h_4]$, where $h_i \in \mathbb{R}^D$. We initially select the feature vector h_1 and concatenate it with all utterances (i.e., h_1, h_2, h_3, and h_4), thereby obtaining the feature representations for the first row, i.e., $t_{(1,1)}, t_{(1,2)}, t_{(1,3)}, t_{(1,4)}$. Subsequently, we take h_2, h_3, and h_4 in sequence and repeat the above procedure. Ultimately, this results in 16 shift samples, with the dimension of the features $t_{(i,j)}$ for each sample being $2D$, where $i = j = 1, 2, 3, 4$.
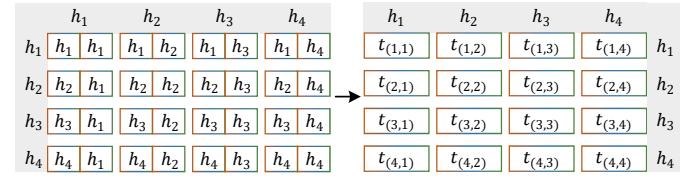
Fig. 3. Example of generating shift features. Here, h_i denotes the feature vector of the utterance u_i.

When the number of utterances in a dialogue (i.e., dialogue length) is excessive, recognizing sentiment shifts between an utterance and long distant utterances becomes meaningless. In other words, intuitively, exploring sentiment shifts between nearby utterances is more meaningful. Moreover, for long dialogues, detecting sentiment shifts between all utterances may consume a substantial amount of computational resources. Therefore, we first evenly divide a dialogue into M/B segments (B is the number of utterances per segment), and then generate sentiment shift labels and corresponding features in each segment.

3) Shift Classification: Similar to the emotion classification, we subject the shift feature representation T to a classifier for shift classification, mathematically formulated as:

$$\begin{aligned} \hat{Z} &= \texttt{softmax}(T\Theta_z + b_z), \\ \hat{o}_{(i,j)} &= \texttt{argmax}(\hat{z}_{(i,j)}[\eta]), \end{aligned} \quad (12)$$

where Θ_z and b_z are learnable parameters for the sentiment shift classifier; $\hat{Z}$ can be denoted as $[\hat{z}_{(1,1)}, \hat{z}_{(1,2)}, \ldots, \hat{z}_{(M,M)}]$, $\hat{z}_{(i,j)}$ represents the shift probability vector constructed based on utterances u_i and u_j, and $\hat{o}_{i,j}$ denotes the corresponding predicted shift label. During the training phase, the cross-entropy loss is employed for network learning:

$$\mathcal{L}_o = -\frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{C_o} z_{(i,j),k} \log \hat{z}_{(i,j),k}, \quad (13)$$

where C_o denotes the number of shift categories ($C_o = 2$ in this work); $z_{(i,j),k}$ signifies the ground-truth probability

TABLE I
MERGING SCHEME OF EMOTION LABELS

Sentiment	IEMOCAP-6	IEMOCAP-4	MELD	CMU-MOSEI
Negative	Sad, Angry, Frustrated	Sad, Angry	Negative	Highly Negative, Negative, Weakly Negative
Neutral	Neutral	Neutral	Neutral	Neutral
Positive	Happy, Excited	Happy	Positive	Weakly Positive, Positive, Highly positive

distribution of the shift sample constructed based on u_i and u_j in the k-th category, and $\hat{z}_{(i,j),k}$ is the predicted probability of the shift sample.

E. Sentiment Classification

Following the emotion classification, we can easily apply GraphSmile to the MSAC task by adding a sentiment classifier. The sentiment classification is represented by the following mathematical formula:

$$\hat{W} = \texttt{softmax}(H\Theta_w + b_w),$$
$$\hat{s}_i = \texttt{argmax}(\hat{w}_i[\phi]). \tag{14}$$

Here, Θ_w and b_w are the learnable parameters specific to the sentiment classifier; $\hat{W}$ can be indicated as $[\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_M]$, and $\hat{w}_i$ represents the sentimental probability vector for utterance u_i, with $\hat{s}_i$ denoting the corresponding predicted sentimental label. The loss function for sentiment classification can be expressed as:

$$\mathcal{L}_s = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{C_s} w_{i,j} \log \hat{w}_{i,j}, \tag{15}$$

where C_s is the number of sentiment categories, $w_{i,j}$ is the ground-truth probability distribution of utterance u_i in the j-th category, and $\hat{w}_{i,j}$ denotes the predicted probability of the utterance.

F. Total Training Objective

The MERC task is jointly learned alongside the MSAC task to simultaneously extract emotional and sentimental cues in the dialogue. Uniting all tasks together, we can derive the total training objective, represented by the following formula:

$$\mathcal{L}_{total} = \mathcal{L}_e + \lambda_s \mathcal{L}_s + \lambda_o \mathcal{L}_o + \beta \|\Theta_{total}\|, \tag{16}$$

where λ_s and λ_o are trade-off parameters balancing these three losses; β is the L2-regularizer weight, and Θ_{total} is the set of all learnable parameters.

IV. EXPERIMENTAL SETUP

A. Baseline

To substantiate the effectiveness of GraphSmile, we select a multitude of baseline models for comparison. These baselines are categorized into graph-based models and non-graph-based models. The former encompasses COGMEN [50], Joyful [61], DER-GCN [62], DCGCN [33], MMGCN [17], MM-DFN [37], M3Net [20]; whereas the latter includes MetaDrop [63], UniMSE [11], EmoLR [64], Dialogue-CRN [65], SACL-LSTM [4].

B. Dataset

We conduct experiments across three datasets, including IEMOCAP, MELD, and CMU-MOSEI. The statistical information of all datasets are shown in Table II. IEMOCAP [47] is a renowned multimodal corpus of dyadic conversations that captures emotional interactions from ten participants in both scripted and improvisational settings. We employ two IEMO-CAP ways, one including all six emotions, i.e., IEMOCAP-6, and the other including four emotions, i.e., IEMOCAP-4, with Excited and Frustrated removed. MELD [48] is an esteemed collection of multimodal multiparty conversations, encompassing 1,433 dialogues from 304 speakers in the television series "Friends", totaling 13,708 utterances. CMU-MOSEI [49] is an extensive multimodal affective collection that aggregates video commentary segments from YouTube, featuring a wide range of 1,000 speakers. To be applicable to GraphSmile, we only use the sentiment annotation in this work. Referring to previous work [50], we divide sentiment intensity into seven distinct classes, encompassing Highly Negative, Negative, Weakly Negative, Neutral, Weakly Positive, Positive, and Highly Positive. The corresponding intensity ranges for these classes are [-3, -2), [-2, -1), [-1, 0), 0, (0, 1], (1, 2], and (2, 3], respectively.

TABLE II
STATISTICAL INFORMATION OF ALL DATASETS

Dataset	Conversation		Utterance	
	train+val	test	train+val	test
IEMOCAP-6	120	31	5,810	1,623
IEMOCAP-4	120	31	3,600	943
MELD	1,153	280	11,098	2,610
CMU-MOSEI	2,549	646	18,198	4,662

The 10% of dialogues from the combination of training and validation sets are selected as the validation set.

C. Training Detail

We utilize accuracy (ACC) and weighted F1 (WF1) as the primary evaluation metrics, with F1 being used as the metric for each individual emotion category. To test the significance of performance improvements, a paired t-test is performed with the default significance level set at 0.05. All experiments are conducted on an NVIDIA GeForce RTX 3090, using CUDA version 11.7 and the deep learning framework PyTorch 2.0.0. AdamW is employed as the optimizer, with the L2 regularization factor of 1e-3. Other hyperparameter

settings are shown in Table III. The code is released at https://github.com/lijfrank-open/GraphSmile.

TABLE III
DESCRIPTION OF PARTIAL HYPERPARAMETERS

Dataset	LR	BS	DR	L	P	B	λ_s	λ_o
IEMOCAP-6	1e-4	16	0.2	7	17	19	1.0	0.7
IEMOCAP-4	3e-4	16	0.2	4	5	10	0.6	0.6
MELD	7e-5	16	0.2	5	3	3	0.5	0.2
CMU-MOSEI	8e-5	32	0.4	2	5	2	0.8	1.0

The abbreviations "LR", "BS", and "DR" denote the learning rate, batch size, and dropout rate, respectively.

V. EXPERIMENTAL RESULT

A. Main Result for Emotion Recognition

The results for the MERC task on all datasets are reported in Table IV. Overall, our proposed GraphSmile outperforms all baseline models on these datasets. Among the reproduced baseline models, M3Net achieves relatively balanced results across all datasets, benefiting from GNNs' thorough exploration of multivariate relationships and multi-frequency signals. The poor performance of DialogueCRN and SACL-LSTM on the CMU-MOSEI dataset indicates that models based on recurrent neural networks struggle to capture the emotional cues present in that dataset. GraphSmile surpasses all graph-based methods on the IEMOCAP (6-way) dataset. For instance, GraphSmile's accuracy is 5.67%, 3.33%, and 1.85% higher than that of MMGCN, MM-DFN, and M3Net on this dataset, respectively. On the IEMOCAP (4-way) dataset, our model improves the weighted F1 score by 2.02% and 0.82% compared to the strong baselines COGMEN and Joyful, respectively. These advanced results are attributed to the significant role played by the GSF module in cross-modal modeling. On the other two more challenging datasets, GraphSmile continues to maintain its advantage. For example, compared to M3Net, GraphSmile's accuracy is 1.11% and 3.15% higher on the MELD and CMU-MOSEI datasets, respectively. In addition to the contribution of the GSF module, the contribution of SDP in sentimental dynamics modeling also contributes to the performance advantage of GraphSmile.

The table also presents the F1 scores for each emotion. In the IEMOCAP (6-way) dataset, compared to other emotions, all models achieve the highest and second-highest scores on Sad and Excited, respectively, and the lowest on Happy. This indicates that Sad and Excited are relatively easier to distinguish than Happy. Relative to other models, our GraphSmile demonstrates a more balanced performance across all emotions, especially achieving the best score on the challenging emotion Happy. As can be seen from the results of the IEMOCAP (4-way) dataset, due to the reduction in the number of categories in the dataset, these models have achieved satisfactory results on all emotions. Among these baselines (with the exception of MMGCN), there is always one or more emotions with significantly lower scores than the others, indicating performance imbalance. In contrast, GraphSmile consistently achieves relatively high results across all emotions. Similar to the results on the IEMOCAP (6-way) dataset, our GraphSmile scores significantly higher on Happy than other models. In the MELD dataset, compared to other emotions, all models perform poorly on Fear and Disgust. The direct cause of this phenomenon is that these two emotions are extreme minority classes in the MELD dataset. Since samples labeled as Neutral constitute the majority of all samples, all models achieve the best F1 scores on Neutral. Compared to other models, SACL-LSTM and GraphSmile achieves the best and second-best scores on the minority class Disgust, respectively. In the CMU-MOSEI dataset, as each utterance may correspond to multiple emotions, we employ seven sentiment labels instead of emotion ones for our experiments. Table IV shows that the classification task for the CMU-MOSEI dataset is quite challenging, with all models struggling to distinguish each category, particularly failing to correctly identify Highly Negative. Compared to the baselines, GraphSmile achieves the best results on Negative, Weakly Positive, and Highly Positive.

B. Main Result for Sentiment Analysis

The proposed GraphSmile not only demonstrates exceptional performance in MERC but also successfully extends to the MSAC task. In Table V, we collect the experimental results of GraphSmile performing the MSAC task across multiple datasets. The results obtained based on the hyperparameter settings in Table IV show that even without task-specific hyperparameter tuning, GraphSmile can still achieve commendable performance, especially surpassing all baselines on the IEMOCAP (4-way) dataset. The reduction in the number of categories leads to a significant performance improvement for all models on datasets other than IEMOCAP (4-way). After hyperparameter optimization specifically for the MSAC task, GraphSmile's performance is further enhanced, outperforming all baselines on most datasets. Although SACL-LSTM achieves the best scores on the IEMOCAP (6-way) dataset, its performance on other datasets is not prominent. Similar to the results of the MERC task, DialogueCRN and SACL-LSTM show relatively weaker capabilities in capturing sentimental cues on the CMU-MOSEI dataset, with their performance significantly lower than other models. Additionally, the table lists the F1 score for each sentiment category in detail. In the IEMOCAP (6-way) dataset, the model scores lower in identifying Neutral, indicating that distinguishing Neutral is more challenging compared to other sentiments. A similar phenomenon is observed in the CMU-MOSEI dataset. In the IEMOCAP (4-way) dataset, the recognition of Negative is the most effective, while in the MELD dataset, the recognition score of Neutral is the highest. These results indicate that there are variations in the recognition difficulty for different datasets and sentiments.

C. Ablation Study

To comprehensively evaluate the functionality of various components in GraphSmile, we report a series of ablation results in Table VI. These studies systematically examine several

TABLE IV
MAIN RESULTS FOR THE MERC TASK ON ALL DATASETS

Method	IEMOCAP-6							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	ACC	WF1
MetaDrop[†]	–	–	–	–	–	–	69.38	69.59
UniMSE[†]	–	–	–	–	–	–	70.56	70.66
COGMEN[†]	51.90	81.70	68.60	66.00	75.30	58.20	68.20	67.60
EmoLR[†]	–	–	–	–	–	–	68.53	68.12
Joyful[†]	<u>60.94</u>	**84.42**	68.24	69.95	73.54	<u>67.55</u>	70.55	71.03
DER-GCN[†]	58.80	79.80	61.50	**72.10**	73.30	**67.80**	69.70	69.40
DCGCN[†]	–	–	–	–	–	–	–	68.31
DialogueCRN[‡]	53.85	82.66	71.03	62.33	<u>77.64</u>	58.81	68.70	68.82
MMGCN[‡]	47.10	81.91	66.44	63.51	76.17	59.06	67.10	66.81
MM-DFN[‡]	43.36	<u>83.23</u>	70.03	70.19	73.11	64.01	69.44	68.83
SACL-LSTM[‡]	51.30	82.25	**71.39**	67.78	75.26	66.94	70.55	70.60
M3Net[‡]	60.93	78.84	70.14	68.06	77.11	67.42	<u>70.92</u>	<u>71.07</u>
GraphSmile	**63.09**	83.16	<u>71.07</u>	<u>71.38</u>	**79.66**	66.84	**72.77**	**72.81**

Method	IEMOCAP-4					
	Happy	Sad	Neutral	Angry	ACC	WF1
COGMEN[†]	–	–	–	–	–	84.50
Joyful[†]	80.10	<u>88.10</u>	<u>85.10</u>	**88.10**	–	<u>85.70</u>
DialogueCRN[‡]	75.64	82.05	81.16	85.22	81.34	81.28
MMGCN[‡]	79.56	79.33	79.27	81.42	79.75	79.72
MM-DFN[‡]	<u>80.30</u>	80.41	79.16	<u>85.64</u>	80.91	80.83
SACL-LSTM[‡]	75.64	78.75	81.83	85.47	80.70	80.74
M3Net[‡]	80.00	85.16	84.63	81.90	<u>83.67</u>	83.57
GraphSmile	**87.94**	**88.67**	85.41	84.73	**86.53**	**86.52**

Method	MELD								
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	ACC	WF1
MetaDrop[†]	–	–	–	–	–	–	–	66.63	<u>66.30</u>
UniMSE[†]	–	–	–	–	–	–	–	65.09	65.51
EmoLR[†]	–	–	–	–	–	–	–	66.69	65.16
Joyful[†]	76.80	51.91	–	41.78	56.89	–	50.71	62.53	61.77
DER-GCN[†]	**80.60**	51.00	10.40	41.50	64.30	10.30	**57.40**	66.80	66.10
DCGCN[†]	–	–	–	–	–	–	–	–	66.25
DialogueCRN[‡]	76.15	56.72	<u>18.67</u>	38.29	63.21	27.69	50.67	62.38	63.32
MMGCN[‡]	78.62	57.78	3.77	40.35	63.60	12.20	<u>53.68</u>	66.02	64.55
MM-DFN[‡]	79.84	58.43	15.79	31.65	64.01	28.04	53.60	<u>67.05</u>	65.21
SACL-LSTM[‡]	77.42	58.50	**20.41**	39.58	62.76	**34.71**	52.08	64.52	64.55
M3Net[‡]	79.14	**59.54**	13.33	**42.86**	**65.05**	21.69	53.54	66.59	65.83
GraphSmile	<u>80.35</u>	<u>59.11</u>	18.18	<u>42.46</u>	<u>64.99</u>	<u>32.43</u>	53.67	**67.70**	**66.71**

Method	CMU-MOSEI								
	HNegative	Negative	WNegative	Neutral	WPositive	Positive	HPositive	ACC	WF1
COGMEN[†]	–	–	–	–	–	–	–	43.90	–
DialogueCRN[‡]	0.00	4.29	7.98	25.09	51.80	3.22	0.00	37.88	26.55
MMGCN[‡]	0.00	<u>19.51</u>	**43.75**	**42.45**	54.64	**36.13**	0.00	<u>45.67</u>	<u>44.11</u>
MM-DFN[‡]	0.00	16.98	37.94	<u>39.64</u>	<u>56.58</u>	32.51	**8.51**	45.29	42.98
SACL-LSTM[‡]	0.00	0.00	0.00	17.87	55.28	0.00	0.00	38.60	25.95
M3Net[‡]	0.00	12.50	37.26	33.29	56.10	33.94	0.00	43.67	41.12
GraphSmile	0.00	**28.79**	<u>43.50</u>	39.19	**57.30**	<u>35.95</u>	6.25	**46.82**	**44.93**

The marker [†] indicates the results (%) from the original paper, and the marker [‡] indicates the replication results (%) based on the open source code and our dataset setting. In DialogueCRN and SACL-LSTM, multimodal features are directly concatenated to meet their input requirements. The abbreviations "HNegative" and "WNegative" denote Highly Negative and Weakly Negative, respectively, and other sentiments by analogy.

TABLE V
MAIN RESULTS FOR THE MSAC TASK ON ALL DATASETS

Method	IEMOCAP-6					IEMOCAP-4				
	Negative	Neutral	Positive	ACC	WF1	Negative	Neutral	Positive	ACC	WF1
DialogueCRN[‡]	87.87	67.09	_89.96_	83.43	83.52	83.94	76.81	79.39	80.70	80.34
MMGCN[‡]	87.07	68.17	88.31	82.75	82.94	84.97	78.80	79.09	81.87	81.56
MM-DFN[‡]	**89.08**	71.92	87.85	84.60	84.69	84.79	76.99	80.95	81.34	81.03
SACL-LSTM[‡]	_88.82_	**72.95**	89.85	**85.15**	**85.35**	85.23	80.23	80.53	82.61	82.47
M3Net[‡]	87.52	68.35	**91.93**	84.29	84.19	87.14	82.61	81.53	84.52	84.44
GraphSmile[†]	86.87	69.06	89.21	83.73	83.66	_87.76_	_82.93_	**87.23**	_85.79_	_85.71_
GraphSmile[♯]	88.28	_72.89_	89.93	_84.97_	_85.09_	**89.45**	**85.53**	_87.02_	**87.49**	**87.48**

Method	MELD					CMU-MOSEI				
	Neutral	Positive	Negative	ACC	WF1	Negative	Neutral	Positive	ACC	WF1
DialogueCRN[‡]	**80.05**	66.67	69.34	_74.18_	_73.96_	20.87	12.71	70.51	54.11	48.50
MMGCN[‡]	79.03	66.60	69.36	73.56	73.46	**58.10**	38.62	78.47	66.66	_65.89_
MM-DFN[‡]	77.89	66.05	**70.83**	73.30	73.27	53.67	36.11	_78.67_	66.35	64.62
SACL-LSTM[‡]	79.08	66.67	69.16	73.56	73.44	10.59	0.00	71.73	55.28	44.50
M3Net[‡]	78.46	_67.12_	69.44	73.33	73.32	53.08	25.92	77.64	64.72	61.66
GraphSmile[†]	79.15	64.84	70.08	73.60	73.40	56.14	_39.80_	78.18	_66.83_	65.69
GraphSmile[♯]	_79.35_	**67.89**	_70.72_	**74.44**	**74.31**	_56.99_	**41.02**	**79.36**	**67.73**	**66.73**

The marker [‡] indicates the replication results (%) based on the open source code and our dataset setting. The marker [†] indicates the result (%) of the MSAC task with the hyperparameters held constant (i.e., the settings in Table III), while the marker [♯] represents the result (%) after optimizing the hyperparameters.

TABLE VI
RESULTS FOR ABLATION STUDIES ON ALL DATASETS

Method	IEMOCAP-6		IEMOCAP-4		MELD		CMU-MOSEI	
	ACC	WF1	ACC	WF1	ACC	WF1	ACC	WF1
GraphSmile[♯]	72.77	72.81	86.53	86.52	67.70	66.71	46.82	44.93
No $\mathcal{L}_s$[♯]	71.53	71.52	85.15	85.15	67.39	66.30	45.08	43.37
No $\mathcal{L}_o$[♯]	71.41	71.42	84.52	84.50	67.62	66.59	45.20	44.03
No $\mathcal{L}_s$ & $\mathcal{L}_o$[♯]	70.67	70.49	84.41	84.40	67.55	66.43	44.29	43.54
No Res[♯]	61.98	61.35	79.64	79.58	52.07	46.24	41.75	40.12
No FC-Res[♯]	70.43	70.39	85.37	85.36	67.36	66.30	45.03	43.28
No Seg[♯]	71.66	71.69	85.47	85.43	67.59	66.59	45.89	44.53
No Textual[♯]	55.70	54.23	62.88	62.22	50.15	42.27	32.98	28.28
No Visual[♯]	71.29	71.23	86.21	86.20	67.39	66.26	45.32	43.57
No Acoustic[♯]	68.45	68.26	83.35	83.28	67.55	66.40	46.56	44.57
GraphSmile[♮]	84.97	85.09	87.49	87.48	74.44	74.31	67.73	66.73
No $\mathcal{L}_e$[♮]	84.41	84.53	85.47	85.40	73.98	73.84	66.59	65.91
No $\mathcal{L}_o$[♮]	84.10	84.18	85.26	85.23	74.18	74.10	66.71	66.20
No $\mathcal{L}_e$ & $\mathcal{L}_o$[♮]	83.30	83.54	85.58	85.55	73.91	73.77	66.54	65.64
No Res[♮]	78.56	78.25	81.02	80.83	58.08	56.61	62.26	61.06
No FC-Res[♮]	84.35	84.38	86.00	85.90	73.72	73.64	66.20	65.69
No Seg[♮]	83.49	83.64	85.37	85.32	74.37	74.22	66.44	65.88
No Textual[♮]	71.04	70.42	66.28	60.76	52.72	52.10	54.95	47.80
No Visual[♮]	84.53	84.69	85.58	85.57	73.83	73.75	66.30	65.81
No Acoustic[♮]	83.24	82.87	84.31	84.25	73.72	73.62	66.47	65.57

The marker [♯] indicates the result (%) of the MERC task, while the marker [♮] indicates the result (%) of the MSAC task. The abbreviation "Res" refers to our designed residual connection, "FC-Res" denotes the fully connected layer in the residual connection, and "Seg" signifies the utterance segmentation strategy when executing shift classification.

key factors, including multitask learning, residual connection, utterance segmentation, and modality configuration.

Multitask Learning: In ablation studies of multitask learning, we design three experiments, i.e., removing the sentiment classification task (No $\mathcal{L}_s$), removing the shift classification task (No $\mathcal{L}_o$), and removing both tasks simultaneously (No $\mathcal{L}_s$ & $\mathcal{L}_o$). The results indicate that the removal of any single task adversely affects the overall performance of the model. In the MERC task, the performance degradation attributable to the excision of $\mathcal{L}_s$ substantiates the capacity of the MSAC task to enhance the MERC task's comprehension of emotions in a dialogue through knowledge complementation. Reciprocally, in the MSAC task, the performance decline induced by the ablation of $\mathcal{L}_e$ evidences that the MERC task can facilitate the MSAC task in understanding sentiments. In both tasks, the decrement in efficacy consequent to the elimination of $\mathcal{L}_o$ corroborates our SDP module can enable the model to capture the sentiment dynamics in a conversation, thus ensuring precise classification in scenarios with abrupt sentimental transitions. In addition, on the IEMOCAP dataset, the removal of $\mathcal{L}_o$ leads to a more significant performance drop compared to the removal of $\mathcal{L}_s$; conversely, on the MELD and CMU-MOSEI datasets, the observed phenomenon is the opposite of that on the IEMOCAP. This could be attributed to the abundance of long dialogues contained in the IEMOCAP dataset, whose sentimental changes is more complex relative to the other two datasets, thus necessitating a more urgent need for modeling sentiment dynamics.

Residual Connection: During evaluating the influence of residual connections on the efficacy, it is observed that the removal of residual connections (No Res) markedly impairs the model's performance across all datasets. These outcomes can be ascribed to several pivotal factors: (1) Compromised modality integration capability. The elimination of residual connections precipitates a direct diminution in the model's capacity to integrate intra-modal contextual and inter-modal associative cues that are propagated in an alternating fashion. (2) Loss of neighboring information. In the absence of residual connections, the model may exhibit a propensity to prioritize long-range information propagation, forgetting the affective cues proffered by neighboring nodes. (3) Over-smoothing predicament. The model devoid of residual connections are susceptible to the over-smoothing dilemma as the network's depth intensifies, leading to a convergence in the feature representations of disparate nodes. (4) Constrained affective expression facility. Simplified graph convolutional layers exhibit constraints in expressing emotions and sentiments due to the lack of non-linear transformation, a deficiency that is amplified by the absence of residual connections. To delve deeper into the role of the fully connected layer in the residual connection, an ablation experiment is conducted by selectively removing the fully connected components of residual connections (No FC-Res). The experimental outcomes indicate a decrement in performance relative to the complete GraphSmile across the four datasets, thereby underscoring the significant role of the fully connected layer in augmenting the model's capacity for affective expression.

Utterance Segmentation: In the shift classification task, we adopt a segmentation strategy for utterances. The relinquishment of the segmentation strategy (No Seg) results in a degradation in the model's performance. This substantiates the efficacy of the strategy in capturing sentiment dynamics, i.e., facilitating the model's perception of sentimental transitions and continuities among proximate utterances, thereby enabling a more nuanced comprehension of useful sentimental shifts and continuities by the model.

Modality Configuration: The empirical outcomes in modality ablation underscores a degradation in the model's holistic performance following the removal of any modality, thereby revealing the indispensable complementarity of multimodal data in the MERC task. Specifically, across all examined datasets, the elimination of the textual modality (No Textual) engenders a markedly pronounced decrement in performance relative to other modalities. This revelation accentuates the dominant role of the textual modality in the MERC task, proffering more critical emotional information. Concurrently, it implies that the visual and acoustic modalities may carry relatively less emotional cues. Such disparities may stem from the noise embedded in visual and acoustic modalities, such as irrelevant background information, ambient sounds, and multiple facial expressions. Further analysis of the experimental results from the IEMOCAP dataset evinces that the exclusion of the visual modality (No Visual) exerts a relatively muted influence on the model's efficacy compared to the excision of the acoustic modality (No Acoustic). This observation posits that in the IEMOCAP dataset, the visual modality contributes less to MERC than the other two modalities. Based on the results from the MELD and CMU-MOSEI datasets, it can be concluded that the contributions of the visual and acoustic modalities to the model's performance on these two datasets are comparable, yet both are inferior to the textual modality.

D. Time and Memory Overhead

In Table VII, we report the running time and memory usage of our model and its baselines for the MERC task across all datasets. Note that all these experiments are conducted on a single NVIDIA GeForce RTX 3090. Relative to other models, MMGCNN and GraphSmile exhibit an absolute advantage in time expenditure on the IEMOCAP dataset, yet MMGCN's time usage on the other two datasets is significantly higher than that of GraphSmile. M3Net and GraphSmile achieve lower time consumption on the MELD dataset, while M3Net's running time on other datasets is inferior to GraphSmile. The running times of DialogueCRN, MM-DFN, and SACL-LSTM across all datasets are unsatisfactory. MMGCN's memory consumption is considerably lower than other models on all datasets, but its overly simplistic model leads to poor performance on most datasets. M3Net and SACL-LSTM have much higher memory consumption on multiple datasets compared to other models, and the memory expenditure of DialogueCRN and MM-DFN is also higher than that of the proposed model. In summary, GraphSmile not only achieves the best accuracy and F1 scores but also balances the costs of time and memory.

TABLE VII
TIME AND MEMORY OVERHEAD FOR THE MERC TASK ON ALL DATASETS

Method	IEMOCAP-6		IEMOCAP-4		MELD		CMU-MOSEI	
	Time	Memory	Time	Memory	Time	Memory	Time	Memory
DialogueCRN[‡]	5.82	2411.95	4.13	992.86	15.58	557.15	37.26	766.55
MMGCN[‡]	1.81	547.13	1.59	251.09	13.97	191.24	29.23	344.52
MM-DFN[‡]	2.46	1939.07	2.03	909.80	22.87	767.48	37.73	680.09
SACL-LSTM[‡]	3.06	3353.43	2.48	1257.00	14.20	2042.30	14.76	1634.85
M3Net[‡]	6.58	12841.27	2.18	3782.62	4.70	761.19	11.12	4549.76
GraphSmile	1.12	1840.89	0.67	509.88	5.21	513.33	7.38	467.13

The marker ‡ indicates the replication results based on the open source code and our dataset setting. The term "Time" refers to the duration (s) required for training and testing in each epoch, and "Memory" refers to the consumption (MB) of allocated and reserved memory.

E. Impact of Network Depth

Early GNNs commonly demonstrate the notorious over-smoothing phenomenon [60]. This issue refers to the gradual homogenization of node features as network depth increases, leading to a sharp decline in model performance after reaching a certain threshold. In Fig. 4, the dark blue and dark green lines intuitively display the performance variation of GraphSmile for the MERC task with network depth. On the IEMOCAP (6-way) dataset, GraphSmile's performance improves with the increase of network layers, slightly declines after reaching the peak, and then maintains a slight fluctuation. The performance changes on the IEMOCAP (4-way) and MELD datasets exhibit a similar pattern. These phenomena indicate that GraphSmile successfully circumvents the over-smoothing issue. In Fig. 4, we also illustrate the performance variation of the model without residual connections (i.e., not employing Eq. 2) adopting light blue and light green lines. It can be observed that as the network depth increases, the performance of this model (i.e., the incomplete GraphSmile) rapidly declines after reaching a certain threshold, showing signs of over-smoothing. The performance comparison between the complete and incomplete versions substantiates the effectiveness of residual connections in alleviating the over-smoothing problem. Despite GraphSmile is less effective against over-smoothing on the CMU-MOSEI dataset due to the higher classification difficulty, the complete model still indicates a significant performance advantage over the incomplete version.

F. Impact of Window Size

During the construction of multimodal dialogue graphs, we introduce two sliding windows to eliminate irrelevant relationships. For ease of experimentation, we employ two equal-length windows, i.e., setting the sizes of the past and future windows to be equal. Fig. 5 details the performance of GraphSmile for the MERC task under different window sizes. Taking the IEMOCAP (6-way) dataset as an example, when the window size is 0 (i.e., without considering contextual information), the performance of GraphSmile drops to the lowest, which is also reflected in the IEMOCAP (4-way) and MELD datasets. These results indicate that context

modeling plays a crucial role in the ERC task. As expected, the performance trends across all datasets exhibit a pattern of initial increase followed by a decrease as the window size increases, eventually fluctuating slightly within a certain range. This suggests that there is an optimal window size that allows the model to capture sufficient contextual information while avoiding the introduction of excessive noise.

G. Embedding Visualization

To visually display the classification efficacy of GraphSmile for the MERC task, we employ T-SNE [66] to visualize the utterance embeddings. As depicted in Fig. 6, GraphSmile effectively segregates samples of distinct emotions in the IEMOCAP dataset, thereby substantiating its potency in emotion classification. Although MMGCN can to some extent differentiate emotional samples, its performance does not match that of our proposed model. Similar to GraphSmile, M3Net achieves commendable classification outcomes on this dataset. Analogous phenomena are observed in the MELD dataset. Furthermore, in contrast to the performance on the IEMOCAP and MELD datasets, GraphSmile exhibits suboptimal results on the CMU-MOSEI dataset, aligning with the outcomes of MMGCN and M3Net.

H. Limitation

Like much of the previous work, GraphSmile has the limitation when dealing with similar emotions. As shown in Fig. 7, on the IEMOCAP (6-way) dataset, Happy samples have a high probability of being identified as the positive emotion Excited. Conversely, Excited samples also have a certain probability of being classified as Happy. Another limitation of GraphSmile is its difficulty in correctly detecting minority class samples on the extremely imbalanced dataset. For instance, on the MELD dataset, samples of Fear and Disgust, which are extreme minority classes, are challenging to classify correctly and tend to be identified as the majority class Neutral. Additionally, visual information (e.g., facial expressions) and acoustic information (e.g., intonation) should theoretically contain significant emotional cues. However, as Table VI reflects, due to the presence of a large amount of
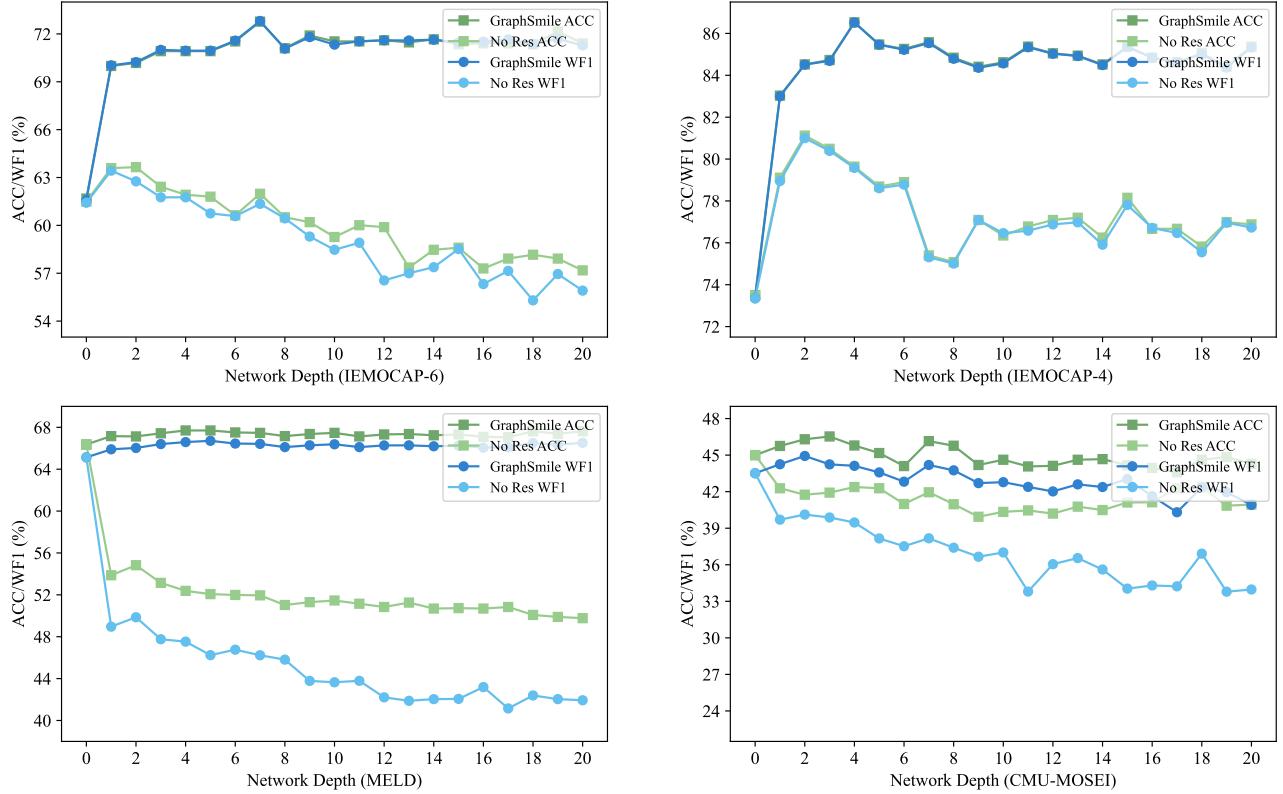
Fig. 4. Performance variations of GraphSmile for the MERC task with the network depth. The abbreviation "Res" indicates the residual connection operation.
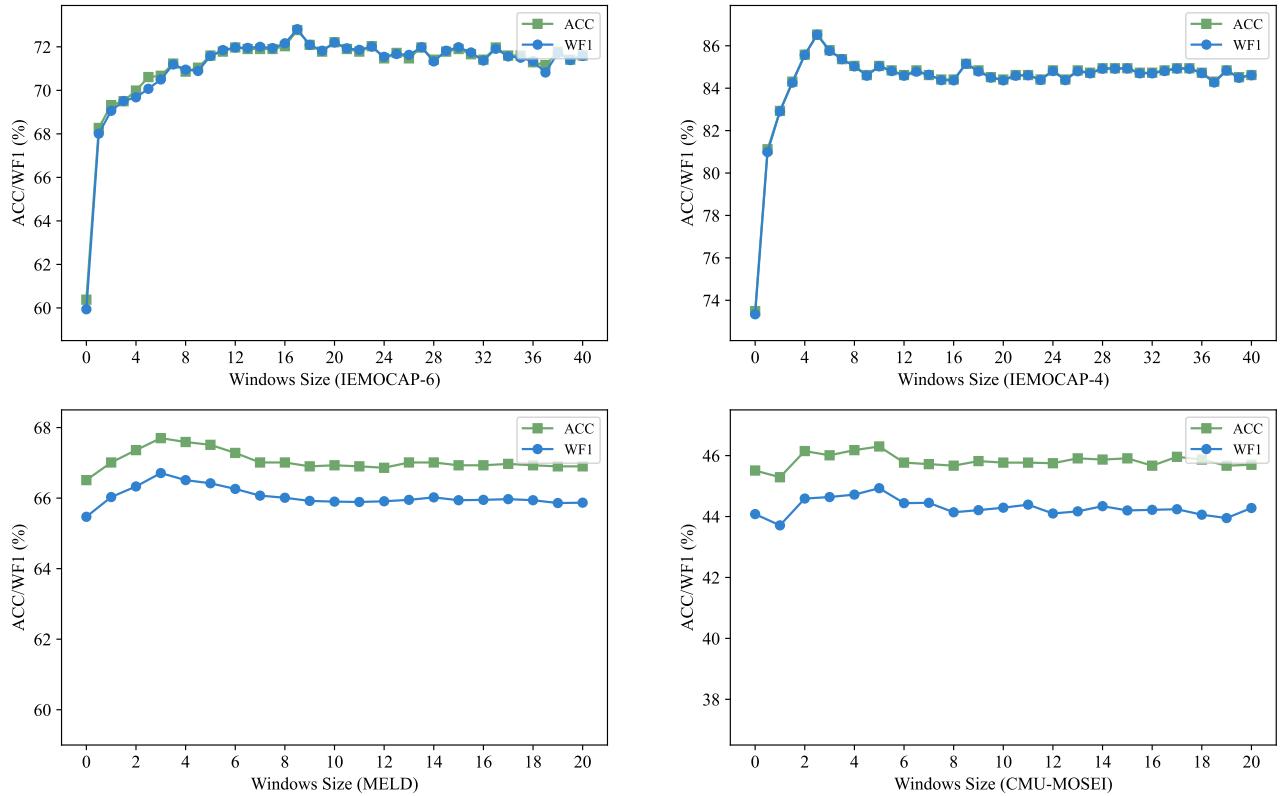

Fig. 5. Performance variations of GraphSmile for the MERC task with the window size. Considering the average dialogue lengths, we set different size ranges for these datasets.

(a) Initial (IEMOCAP-6)   (b) MMGCN (IEMOCAP-6)   (c) M3Net (IEMOCAP-6)   (d) GraphSmile (IEMOCAP-6)

(e) Initial (IEMOCAP-4)   (f) MMGCN (IEMOCAP-4)   (g) M3Net (IEMOCAP-4)   (h) GraphSmile (IEMOCAP-4)

(i) Initial (MELD)   (j) MMGCN (MELD)   (k) M3Net (MELD)   (l) GraphSmile (MELD)

(m) Initial (CMU-MOSEI)   (n) MMGCN (CMU-MOSEI)   (o) M3Net (CMU-MOSEI)   (p) GraphSmile (CMU-MOSEI)
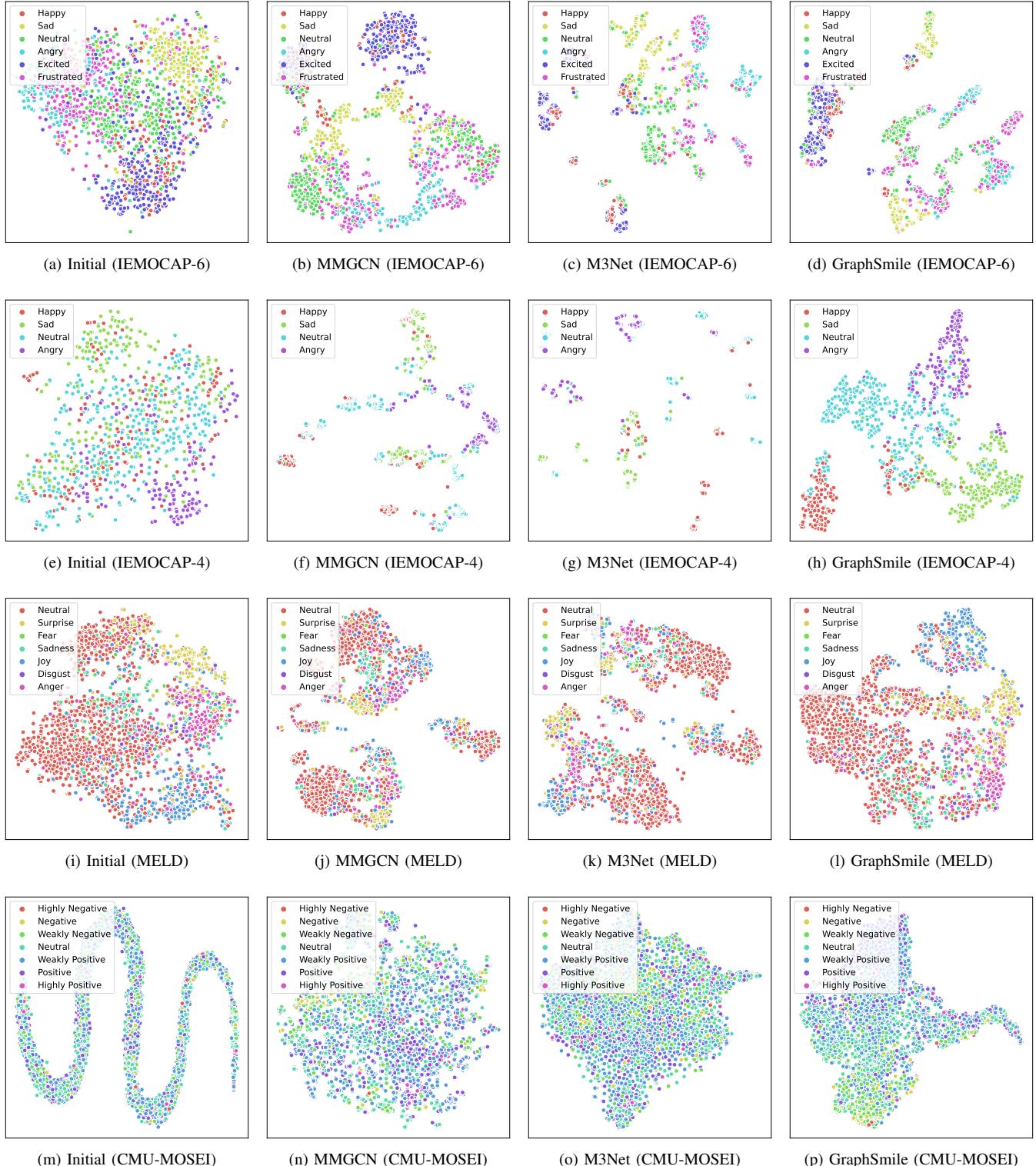
Fig. 6. Visualization of utterance embeddings for all datasets. Here, "Initial (IEMOCAP-6)" signifies the t-SNE visualization of the initial features in the IEMOCAP (6-way) dataset, and "MMGCN (IEMOCAP-6)" indicates the visualization of the embeddings extracted by MMGCN in the same dataset. Similar nomenclature is applied to other figures.
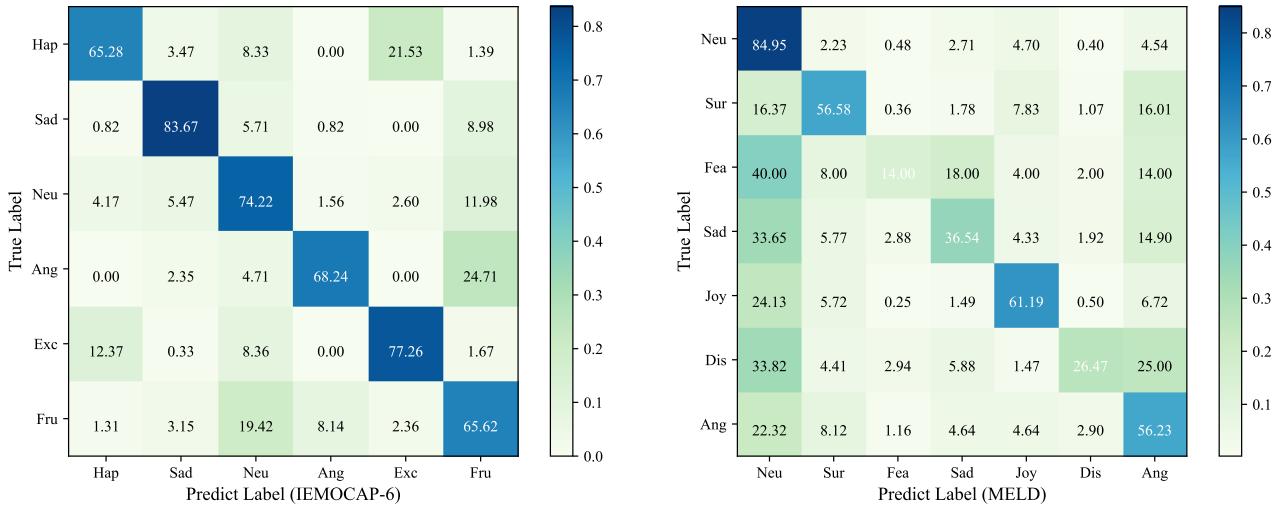
Fig. 7. Confusion matrices on the IEMOCAP (6-way) and MELD datasets. The abbreviation "Hap" denotes Happy and other emotions by analogy.

noise, they do not leverage their unique advantages in dialogue affective tasks. This may lead GraphSmile to favor the textual modality in cross-modal modeling, thereby reducing the effective mining of visual and acoustic information.

VI. CONCLUSION

In this research, we propose an innovative multitask multimodal affective model named GraphSmile, specifically tailored for MERC and MSAC tasks. The proposed model achieves synergistic optimization of these tasks by integrating emotion recognition, sentiment analysis, and sentiment shift detection, thereby delving into the intricate emotional and sentimental cues embedded in multimodal dialogues. GraphSmile employs a unique graph construction manner that not only facilitates comprehensive capture of cross-modal interaction cues but also effectively mitigates redundant connections in the multimodal dialogue graph. To circumvent potential conflicts arising from the fusion of multi-source data, we design a graph structure based multimodal fusion module, termed GSF. This module adeptly harnesses the graph structure, optimizing the integration process by alternately absorbing inter-modal and intra-modal affective dependencies in a layered fashion. In the GSF module, we introduce a residual connection mechanism integrated with fully connected layers, which not only alleviates the over-smoothing issue common in GNNs but also significantly enhances the model's capacity for affective expression. Furthermore, to bolster the model's capability to capture the sentiment dynamics, we develop a plug-and-play auxiliary module known as SDP. This module explicitly models the sentimental transitions between utterances, augmenting the model's discernment of sentimental variances. We have conducted extensive validation experiments on several public benchmark datasets, and the results robustly substantiate the efficacy of GraphSmile and its components.

In future work, we intend to address other salient challenges in the realm of affective computing in conversation. For instance, (1) tackling the differentiation of similar emotions and issues of class imbalance through contrastive learning strategies, and (2) delving deeper into the sentimental dynamics in conversation by considering the temporal logic of transitions and refining the types of shift labels. We will also explore more sophisticated multimodal fusion strategies to capitalize on the unique strengths of each modality and enhance the model's robustness against noise and irrelevant information. Recently, large language models (LLMs) have demonstrated formidable reasoning capabilities, thus integrating MERC and MSAC with LLM technologies will further propel the development of affective community.

REFERENCES

[1] W. X. Jiao, H. Q. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 397–406.

[2] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: Commonsense knowledge for emotion identification in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020, pp. 2470–2481.

[3] D. Ong, J. Su, B. Chen, A. T. Luu, A. Narendranath, Y. Li, S. Sun, Y. Lin, and H. Wang, "Is discourse role important for emotion recognition in conversation?" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 121–11 129, Jun. 2022.

[4] D. Hu, Y. Bao, L. Wei, W. Zhou, and S. Hu, "Supervised adversarial contrastive learning for emotion recognition in conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 835–10 852.

[5] W. Z. Shen, S. Y. Wu, Y. Y. Yang, and X. J. 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*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1551–1560.

[6] W. Zhao, Y. Zhao, and X. Lu, "CauAIN: Causal aware interaction network for emotion recognition in conversations," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 4524–4530, main Track.

[7] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[8] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[9] 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. 01, 2019, pp. 6818–6825.

[10] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, 2022, pp. 1610–1618.

[11] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 7837–7851.

[12] B. Li, H. Fei, L. Liao, Y. Zhao, C. Teng, T.-S. Chua, D. Ji, and F. Li, "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 5923–5934.

[13] Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li, "DialogueTRM: Exploring multi-modal emotional dynamics in a conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, 2021, pp. 2694–2704.

[14] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[15] X. Zhang and Y. Li, "A cross-modality context fusion and semantic refinement network for emotion recognition in conversation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 13 099–13 110.

[16] W. Zheng, J. Yu, R. Xia, and S. Wang, "A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 15 445–15 459.

[17] 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 Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021, Conference Proceedings, pp. 5666–5675.

[18] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 77–89, 2024.

[19] D. Zhang, F. Chen, J. Chang, X. Chen, and Q. Tian, "Structure aware multi-graph network for multi-modal emotion recognition in conversations," *IEEE Transactions on Multimedia*, vol. 26, pp. 3987–3997, 2024.

[20] F. Chen, J. Shao, S. Zhu, and H. T. Shen, "Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 761–10 770.

[21] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, apr 2024, just Accepted.

[22] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Computing Surveys*, vol. 56, no. 9, apr 2024.

[23] S. Li, H. Yan, and X. Qiu, "Contrast and generation make BART a good dialogue emotion recognizer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 002–11 010, Jun. 2022.

[24] X. Song, L. Huang, H. Xue, and S. Hu, "Supervised prototypical contrastive learning for emotion recognition in conversation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5197–5206.

[25] G. Tu, B. Liang, R. Mao, M. Yang, and R. Xu, "Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations," in *Findings of the Association for Com-*

[26] T. Zhang, Z. Chen, M. Zhong, and T. Qian, "Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI '23, 2023.

[27] F. Yu, J. Guo, Z. Wu, and X. Dai, "Emotion-anchored contrastive learning framework for emotion recognition in conversation," in *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4521–4534.

[28] 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)*, 2019, pp. 154–164.

[29] K. Wang, W. Z. Shen, Y. Y. Yang, X. J. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3229–3238.

[30] B. Lee and Y. S. Choi, "Graph based network with contextualized representations of turns in dialogue," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 443–455.

[31] M. Ren, X. Huang, W. Li, D. Song, and W. Nie, "LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 4422–4432, 2022.

[32] D. Zhang, F. Chen, and X. Chen, "DualGATs: Dual graph attention networks for emotion recognition in conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 7395–7408.

[33] Z. Yang, X. Li, Y. Cheng, T. Zhang, and X. Wang, "Emotion recognition in conversation based on a dynamic complementary graph convolutional network," *IEEE Transactions on Affective Computing*, pp. 1–14, 2024.

[34] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "GCNet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8419–8432, 2023.

[35] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNet: Multi-modal fusion network for emotion recognition in conversation," in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 4651–4660.

[36] S. Zou, X. Huang, X. Shen, and H. Liu, "Improving multimodal fusion with main modal transformer for emotion recognition in conversation," *Knowledge-Based Systems*, vol. 258, p. 109978, 2022.

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

[38] G. Tu, T. Xie, B. Liang, H. Wang, and R. Xu, "Adaptive graph learning for multimodal conversational emotion detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19 089–19 097, Mar. 2024.

[39] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 481–492.

[40] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9180–9192.

[41] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 309–325, 2024.

[42] C. Fan, K. Zhu, J. Tao, G. Yi, J. Xue, and Z. Lv, "Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, pp. 1–17, 2024.

putational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 054–14 067.

[43] S. Verma, J. Wang, Z. Ge, R. Shen, F. Jin, Y. Wang, F. Chen, and W. Liu, "Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis," in *2020 IEEE International Conference on Data Mining*, 2020, pp. 561–570.

[44] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569.

[45] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 1009–1021.

[46] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, and R. Xu, "Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis," in *Proceedings of the 29th International Conference on Computational Linguistics*, vol. 29, no. 1. Association for Computational Linguistics, October 2022, pp. 7124–7135.

[47] 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.

[48] 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*, Florence, Italy, 2019, pp. 527–536.

[49] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246.

[50] 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.

[51] 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," *arXiv preprint arXiv:1907.11692*, pp. 1–15, 2019.

[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[53] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, pp. 279–283.

[54] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 59–66.

[55] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[56] 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.

[57] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18 – 24.

[58] 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*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 154–164.

[59] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 6861–6871.

[60] G. Li, M. Müller, G. Qian, I. C. Delgadillo, A. Abualshour, A. Thabet, and B. Ghanem, "DeepGCNs: Making gcns go as deep as cnns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6923–6939, jun 2023.

[61] D. Li, Y. Wang, K. Funakoshi, and M. Okumura, "Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16 051–16 069.

[62] W. Ai, Y. Shou, T. Meng, and K. Li, "DER-GCN: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.

[63] F. Chen, Z. Sun, D. Ouyang, X. Liu, and J. Shao, "Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, 2021, pp. 1064–1073.

[64] J. Wen, G. Tu, R. Li, D. Jiang, and W. Zhu, "Learning more from mixed emotions: A label refinement method for emotion recognition in conversations," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1485–1499, 12 2023.

[65] D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Online: Association for Computational Linguistics, Aug. 2021, pp. 7042–7052.

[66] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.