Learning Dynamic Frame Semantics for Multimodal Sentiment Analysis

Jian Zhuang, Tiantian He, *Member, IEEE*, Yaqing Hou, Yew-Soon Ong, *Fellow, IEEE*, Qiang Zhang, *Senior Member, IEEE*

***Abstract*—In this paper, we investigate how the mathematization of frame semantics can enhance multimodal sentiment analysis (MSA). Existing deep learning approaches to MSA typically focus on single semantic frames. In contrast, frame semantics, a well-recognized theory from linguistics, posits that language cognition and interpretation essentially rely on the coherence of multiple frames divided according to lexical units and encapsulated with diverse background knowledge and conceptual structures. To facilitate MSA with frame semantics, we propose a new model, namely Dynamic-Aware Frame Semantic Analysis (DAFSA). DAFSA first models multimodal sequential features (e.g., text, audio, and vision) as multimodal semantic frames. Then, DAFSA can extract the representations that capture the subtle emotional expressions by dynamically learning and summarizing the coherent structures from intra- and cross-modal frames. Experimental results demonstrate that DAFSA significantly outperforms state-of-the-art approaches to MSA and foundation models whose size is larger than DAFSA up to 100×, showcasing the effectiveness of modeling frame semantics in MSA.**

***Index Terms*—Multimodal sentiment analysis, frame semantics, mathematical modeling.**

I. Introduction

EMOTION is one of the most fundamental physiological responses of humans, influencing thoughts, behaviors, and decision-making. Recently, with the popularization of social media and streaming platforms, people have become increasingly inclined to express their attitudes through videos [1]. Consequently, analyzing the emotional information conveyed in multimodal data has become increasingly important [2]. This process, commonly referred to as multimodal sentiment analysis (MSA), aims to extract and integrate information from multiple modalities, typically including language (textual content), audio (acoustic features), and vision (facial expressions) to infer expressed sentiments [3], [4]. MSA is of great significance and has broad applications, such as human-computer interaction [5], [6], social media analysis [7], [8], and smart justice [9], [10].

Unlike unimodal sentiment analysis [11], [12], MSA requires leveraging different information processing techniques to extract meaningful sentiment information from various modalities. At the same time, it is essential to integrate information from multiple modalities to enable a comprehensive analysis of human emotions. For example, when a speaker sarcastically says, "This movie is really special," relying solely on textual information is insufficient to fully interpret the speaker's emotional state. In this case, it is necessary to consider additional cues such as tone and intonation to comprehensively analyze the speaker's true sentiment from multiple perspectives. Therefore, the efficient integration and utilization of multimodal information plays a crucial role in improving the accuracy of sentiment analysis [13]–[15].

Existing approaches to MSA primarily focus on modeling dynamic interactions within and across modalities. Generally, these methods can be categorized into two types: word-level methods [16], [17] and utterance-level methods [18], [19]. Word-level approaches fuse multimodal features at the word level to capture fine-grained interactions, whereas utterance-level methods focus on global semantic structures by aggregating information at the sentence level. The latter often yields more holistic representations but may overlook subtle word-level sentiment cues.

Despite the impressive performance of these approaches to diverse MSA tasks, most of them overlook the co-occurrence mechanism of multiple semantic frames within a single utterance, which is crucial for the interpretation and cognition of sentiment. For instance, in the statement "The special effects are okay—that is, the storyline is too dull." different lexical units trigger distinct semantic frames that jointly shape the overall sentiment. The phrase "special effects" with "okay" evokes a mildly positive evaluative frame, while "storyline" paired with "too dull" activates a contrasting negative frame

This work was supported in part by the National Key R&D Program of China under Grant 2024YFA1012700, the Liaoning Provincial Central-Guided Local Sci-Tech Development Program under Grant 2025JH6/101000005, the Dalian Science and Technology Innovation Fund under Grant 2024JJ12GX020, the Liaoning Provincial Natural Science Foundation Program under Grant 2024-MSBA-05, the Liaoning Provincial Science and Technology Program under Grant 2024JH2/102600046, the 111 Project under Grant D23006, the Aviation Transformation Programme-Airport (ATP_Airport) of Singapore (Award No. ATP_AIRPORT/2026/ARES/FLASH), and the National Research Foundation, Singapore under its NRF AI-for-Science (AI4S) Challenge Grant (Award NRF-AI4SCH-2025-0007). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. *(Corresponding author: Yaqing Hou, Qiang Zhang.)*

Jian Zhuang, Yaqing Hou and Qiang Zhang are with the College of Computer Science and Technology, Dalian University of Technology (DLUT), China 116024, and are also with the Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, China (e-mail: zhuangj@mail.dlut.edu.cn; houyq@dlut.edu.cn; zhangq@dlut.edu.cn)

Tiantian He is with the Center for Frontier AI Research, Institute of High Performance Computing, Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: he_tiantian@a-star.edu.sg).

Yew-Soon Ong is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, and also with the Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: asysong@ntu.edu.sg).

The source code is available on https://github.com/dutzj/DAFSA.

centered on narrative quality. Moreover, the connective "that is" invokes a contrastive frame that highlights the shift in evaluative focus. The interaction and progression among these frames reveal subtle sentiment transitions that most existing methods [18]–[20] fail to capture.

In this paper, we hypothesize that appropriate mathematization of frame semantics [21], which has been recognized as a famed theory in linguistics, may significantly advance the performance of MSA. Frame semantics suggests that the meaning of language relies on some fundamental structures or scenarios, denoted as "frames". The meaning of words is interpreted by concurrently triggering these frames, rather than by the words themselves. Frames encompass the background and context that should be analyzed to reveal the meaning of words in particular situations. Thus, effective MSA is primarily influenced by accurately capturing the implicit background and context regarding sentiment from the frames that are built with multimodal data.

To this end, we present a novel model, named Dynamic-Aware Frame Semantic Analysis (DAFSA). Unlike existing MSA methods that primarily organize multimodal information through feature fusion, cross-modal interaction, or graph-based relation modeling, DAFSA introduces a structured module that dynamically activates contextually relevant frames during the learning process. This mechanism enables the model to learn the evocation relationships between lexical units and semantic frames, and to identify the affective semantics carried by frame elements within multimodal contexts. As a result, the model constructs multiple frame structures and captures sentiment transformation mechanisms—such as contrast and progression—across frames, thereby enhancing the accuracy and expressiveness of sentiment reasoning.

To develop DAFSA, we achieve the following technical contributions. First, we build a unimodal encoder to capture the temporal dependencies within each modality and generate lexical representations to support the subsequent analysis of frame semantics. Then, we propose the dynamic-aware frame semantic analysis module to model multimodal data as semantic frame structures. This module dynamically learns the structure and features of these frames and integrates these features into lexical units to enhance the model's ability to perceive subtle sentiments. Lastly, we present the Semantic Fusion module, which eliminates redundant semantic frames and fuses lexical representations with frame-based features to form a comprehensive multimodal representation, improving the accuracy and robustness of sentiment analysis.

The main contributions of our work are summarized as follows:

- To the best of our knowledge, we provide the first mathematical formulation of core frame-semantic concepts for MSA, where multimodal sequential features are represented as lexical units and organized into intra-modal and inter-modal semantic frames. This formulation establishes a frame-semantic organization mechanism for multimodal sentiment reasoning.
- We propose Dynamic-Aware Frame Semantic Analysis (DAFSA), which models lexical-unit evocation, frame structure refinement, and frame-level semantic integration through frame learning, frame integration, dynamic-aware selection, and semantic fusion. These designs enable the model to capture multiple semantic frames and their contextual contributions to sentiment reasoning.
- We conduct extensive experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS to evaluate DAFSA from multiple perspectives, including overall effectiveness, component necessity, parameter robustness, computational efficiency, and interpretability. The results demonstrate that DAFSA achieves competitive or superior performance compared with representative MSA baselines and text-only prompted LLMs, with up to approximately $100\times$ fewer parameters than the compared LLMs.

II. Related Work

As one of the most popular topics in natural language processing, multimodal sentiment analysis (MSA) has been extensively explored in recent years [22], [23]. Early methods typically relied on feature concatenation or manual alignment [24], [25], where static features from different modalities were combined and fed into a classifier. While straightforward, these approaches lacked the ability to capture dynamic dependencies and semantic interactions across modalities. To address this, tensor-based fusion methods [18], [19] were proposed to represent multiple modalities through structured combinations, though they often suffer from high computational cost and limited scalability. RNN-based fusion methods [20], [26] were later introduced to model temporal dependencies within and across modalities, improving the representation of emotional dynamics. However, they remain limited in capturing long-range dependencies. In response, attention-based fusion methods [27], [28] enable the model to dynamically focus on emotionally salient segments, enhancing alignment and semantic integration. More recently, graph neural network (GNN)-based fusion methods [29]–[31] have been explored from a non-Euclidean perspective. By organizing multimodal data into graph structures, these methods better capture complex modality relationships and improve the flexibility of fusion strategies.

Although these studies perform excellently in modeling intra-modal and inter-modal interactions, most existing methods [30]–[32] rely on a single semantic frame for sentiment modeling, ignoring the subtle emotional expressions implied by multiple semantic frames.

To address this issue, we introduce frame semantics [21] to systematically analyze sentiment expression and propose an innovative approach called Dynamic-Aware Frame Semantic Analysis (DAFSA). This method can learn multiple semantic frames and structurally models the multiple contextual factors in sentiment expression. We are the first to achieve the mathematical formalization of frame semantics and integrate its core concepts from cognitive linguistics and semantics into the MSA. By formalizing frame semantics, we provide a more structured knowledge representation for sentiment analysis, enabling the model to dynamically activate the relevant semantic frames during sentiment reasoning, thereby enhancing its ability to understand sentiment expressions.

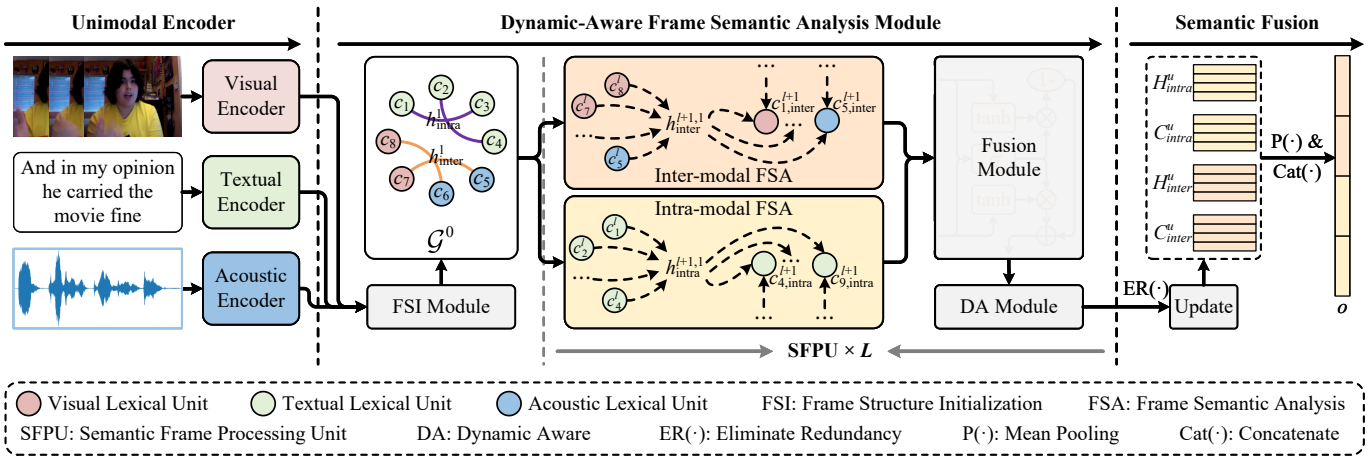

Fig. 1. The overall architecture of the proposed DAFSA framework. The model consists of three main components: (1) Unimodal Encoder: encodes text, audio, and visual modalities into sequential representations as the input basis for subsequent semantic analysis; (2) Dynamic-Aware Frame Semantic Analysis Module: organizes multimodal representations into semantic frames, performs intra- and inter-modal frame semantic learning and fusion, and adaptively updates frame structures through the dynamic-aware mechanism; (3) Semantic Fusion Module: fuses lexical units and semantic frame information into a unified multimodal joint representation for final sentiment reasoning and prediction.

Overall, existing methods generally represent textual, acoustic, and visual signals as sequential features, interaction nodes, or relational structures, and improve sentiment prediction by strengthening fusion, alignment, or dependency modeling among them. In contrast, DAFSA takes lexical units and the semantic frames they evoke as the basic modeling units, and explicitly formalizes how frame structures are initialized, activated, updated, and integrated for sentiment reasoning. This shifts the modeling focus from feature-level interaction to frame-semantic organization, providing a theoretically grounded and computationally executable formulation for MSA.

III. Method

In this section, we conduct a systematic mathematical modeling of frame semantics and accordingly propose the Dynamic-Aware Frame Semantic Analysis (DAFSA) method, enabling its computational representation and effective application in multimodal sentiment analysis tasks. The problem formulation of MSA and necessary notations are first introduced. Then, we elaborate on the three core components of DAFSA: the unimodal encoder constructs lexical unit representations, the dynamic-aware (DA) module performs semantic frame activation and information integration, and the semantic fusion module achieves role fusion and unified multimodal joint representation. The overall architecture of DAFSA is shown in Fig. 1.

A. Problem Definition and Notations

Given a sample with its corresponding sentiment intensity label y, the sample consists of three modalities: text U_t, audio U_a, and vision U_v. Each modality is represented as $U_m = \{u_m^1, u_m^2, \ldots, u_m^{T_m}\} \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{t, a, v\}$, T_m denotes the temporal length of the features, and d_m represents the feature dimension. The objective of the MSA task is to comprehensively model the features of these three modalities to learn a multimodal joint representation, and ultimately to predict the sentiment intensity of the sample. Furthermore, the main notations and their descriptions in this paper are shown in Table I.

TABLE I
SUMMARY OF NOTATIONS AND DESCRIPTIONS, WHERE $m \in \{t, a, v\}$ REPRESENTS TEXT, AUDIO, AND VISUAL MODALITIES, RESPECTIVELY.

Symbol	Dimensionality	Description
U_m	$\mathbb{R}^{T_m \times d_m}$	Sequential features of modality m
X_m	$\mathbb{R}^{T_m \times d}$	Sequential features after unimodal encoder
C	$\mathbb{R}^{N_T \times d}$	Lexical unit feature matrix
E_*	$\mathbb{R}^{N_T \times N_T}$	Activation matrix for semantic frames
H_*	$\mathbb{R}^{T_m \times d}$	Features of semantic frames
E_*^o	$\mathbb{R}^{N_* \times d}$	Set of deduplicated semantic frames
H_*^o	$\mathbb{R}^{N_* \times d}$	Features of deduplicated semantic frames
H_*^u	$\mathbb{R}^{N_* \times d}$	The updated H_*^o in Semantic Fusion module
C_*^u	$\mathbb{R}^{N_T \times d}$	Lexical unit feature after integrating H_*^o
$h_*^{l,i}$	$\mathbb{R}^d$	Representation of the i-th frame at layer l
c_j^l	$\mathbb{R}^d$	Representation of the j-th lexical at layer l
$h_*^{u,i}$	$\mathbb{R}^d$	The i-th semantic frames feature in H_*^u
$c_*^{u,j}$	$\mathbb{R}^d$	The j-th Lexical unit feature in C_*^u
o	$\mathbb{R}^{4d}$	Final multimodal joint representation

Note: $* \in \{\text{intra}, \text{inter}\}$ indicates intra-modal and inter-modal, and l denotes the layer number of SFPU. Uppercase italic letters denote feature matrices, whereas lowercase italic letters denote feature vectors.

B. Unimodal Encoder

To effectively support the subsequent modeling of frame semantics, particularly the construction of lexical unit representations, we first perform contextual encoding of multimodal data. Due to the inherent temporal dependencies in multimodal sequences, processing each time step independently may lead to the loss of temporal information. To address this issue, we adopt modality-specific Bidirectional LSTMs (Bi-LSTMs) [33] to capture contextual dependencies in both

forward and backward directions. The resulting hidden states are then mapped into a unified dimensional space through a fully connected layer, yielding the final unimodal feature representations X_m.

$$X_m = W_m \left(\textit{Bi-LSTM} \left(U_m, \theta_m^{lstm} \right) \right) + b_m, \tag{1}$$

where $X_m \in \mathbb{R}^{T_m \times d}$ represents the output features of the unimodal encoder, and d is the unified feature dimension. θ_m^{lstm} denotes the trainable parameters of the Bi-LSTM. W_m and b_m represent the weight matrix and bias term of the fully connected layer, respectively.

C. Dynamic-Aware Frame Semantic Analysis Network

Language understanding relies on the activation of semantic frames by lexical units and the integration of contextual and semantic information from these frames. To implement this idea, we propose a dynamic-aware frame semantic analysis module (as shown in Fig. 1). We first perform semantic frame structure initialization to establish the associations between lexical units and frames. Then, we conduct semantic frame analysis to further extract semantic information and integrate sentiment features.

1) Frame Structure Initialization Module: The initialization of frame structures essentially serves as a prior specification of possible semantic scenarios or conceptual configurations, providing a candidate semantic space for subsequent lexical activation. In frame semantics, a lexical unit may evoke multiple frames, and a single frame can also be activated by multiple lexical units. Therefore, in MSA, we adopt a fully connected strategy to initialize the multimodal semantic frame structures, ensuring comprehensive coverage of potential semantic relationships in the data.

Lexical unit initialization: Since X_m contains the contextual dependency information of modal features and temporal correlations, we define the unified lexical unit representation as $C = [X_t||X_a||X_v] = \{c_1, c_2, \ldots, c_{N_T}\} \in \mathbb{R}^{N_T \times d}$, where $N_T = T_t + T_a + T_v$ denotes the total number of lexical units across the three modalities, and $||$ denotes concatenation.

Semantic frame initialization: Frame initialization consists of two components, i.e., frame feature initialization and activation matrix initialization. In MSA, considering the heterogeneity among different modalities, for each lexical unit c_i, we construct two types of semantic frames—an intra-modal frame and an inter-modal frame—to capture both modality-specific and cross-modal relationships. Both frames naturally inherit the modality attribute of each lexical unit c_i; the intra-modal frame is associated with lexical units from the same modality, whereas the inter-modal frame is associated with lexical units from different modalities. The features of these frames are initialized by applying the corresponding modality-specific mapping functions to the lexical units $c_i \in \mathbb{R}^d$:

$$\begin{aligned} h_{\text{intra}}^i &= f_{\text{intra}}\left(c_i\right), \\ h_{\text{inter}}^i &= f_{\text{inter}}\left(c_i\right), \end{aligned} \tag{2}$$

where $h_{\text{intra}}^i \in \mathbb{R}^d$ and $h_{\text{inter}}^i \in \mathbb{R}^d$ represent the features of intra- and inter-modal frames, respectively, while $f_{\text{intra}}(\cdot)$ and $f_{\text{inter}}(\cdot)$ are the mapping functions. $H_{\text{intra}} = \{h_{\text{intra}}^1, h_{\text{intra}}^2, \ldots, h_{\text{intra}}^{N_T}\} \in \mathbb{R}^{N_T \times d}$ and $H_{\text{inter}} = \{h_{\text{inter}}^1, h_{\text{inter}}^2, \ldots, h_{\text{inter}}^{N_T}\} \in \mathbb{R}^{N_T \times d}$ denote the intra-modal and inter-modal frame sets respectively.

The activation matrix represents the relationship between lexical units and semantic frames. It is denoted as $E_* \in \mathbb{R}^{N_T \times N_T}$, where $* \in \{\text{intra}, \text{inter}\}$. Specifically, each element of the activation matrix $E_*[i,j]$ is defined as follows:

$$E_*[i,j] = \begin{cases} 1, & \text{if lexical unit } j \text{ in semantic frame } i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Finally, we use $\mathcal{G}^0 = (C, E_{\text{intra}}, E_{\text{inter}}, H_{\text{intra}}, H_{\text{inter}})$ to represent the frame structure. Fig. 1 illustrates the constructed frame structure using lexical unit c_1 as an example, where the yellow and purple lines represent intra- and inter-modal semantic frames, respectively.

2) Semantic Frame Processing Unit: In frame semantics, language understanding relies on the activation of frames by lexical units, and word meaning is interpreted by integrating information from these frames. To implement this process in MSA, we present the Semantic Frame Processing Unit (SFPU). SFPU activates relevant semantic frames through lexical units, learns the structural and semantic properties of these frames, and integrates frame information into lexical features, thereby improving the model's ability to understand sentiment. As shown in Fig. 1, each SFPU includes four components: intra-modal frame semantic analysis (FSA), inter-modal FSA, a fusion module, and a DA module.

Intra-modal FSA: Intra-modal FSA activates intra-modal semantic frames through lexical units to model intra-modal sentiment knowledge, and then integrates the learned frame information into lexical representations, thereby enhancing the model's ability to perceive intra-modal sentiment expressions. This process consists of two steps: Intra-modal Semantic Frame Learning (SFL) and Intra-modal Semantic Frame Integration (SFI).

For Intra-modal SFL, let $h_{\text{intra}}^{l,i}$ and c_j^l represent the feature representation of the i-th intra-modal semantic frame and the representation of the j-th lexical unit at layer l, respectively. Note that $h_{\text{intra}}^{0,i} = h_{\text{intra}}^i$ and $c_j^0 = c_j$. We compute the correlation coefficient between the semantic frame and its associated lexical units, and normalize it to obtain the attention coefficient $\alpha_{ij,\text{intra}}^l$:

$$\alpha_{ij,\text{intra}}^l = \text{Softmax}(\text{LReLU}((a_{\text{intra}}^{l,h})^T W_{\text{intra}}^l [h_{\text{intra}}^{l,i} || c_j^l])), \tag{4}$$

where $\text{LReLU}(\cdot)$ denotes the LeakyReLU activation function, $a_{\text{intra}}^{l,h} \in \mathbb{R}^{2d}$ is a learnable attention vector. The normalized attention coefficients $\alpha_{ij,\text{intra}}^l$ are then used to perform a weighted summation of lexical unit information to update the intra-modal semantic frame features:

$$h_{\text{intra}}^{l+1,i} = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij,\text{intra}}^l W_{\text{intra}}^l c_j^l\Big), \tag{5}$$

where $\sigma(\cdot)$ is a nonlinear activation function.

For Intra-modal SFI, to integrate intra-modal semantic frames, we assess the correlation between each lexical unit

and its associated semantic frame and apply a normalization to obtain the intra-modal attention coefficients $\beta^l_{ji,\text{intra}}$:

$$\beta^l_{ji,\text{intra}} = \text{Softmax}(\text{LReLU}((a^{l,c}_{\text{intra}})^T W^l_{\text{intra}}[c^l_j || h^{l+1,i}_{\text{intra}}])), \quad (6)$$

where $\beta^l_{ji,\text{intra}}$ represents the correlation coefficient between lexical unit j and semantic frame i, $a^{l,c}_{\text{intra}} \in \mathbb{R}^{2d}$ is a learnable attention vector. The normalized attention coefficients $\beta^l_{ji,\text{intra}}$ are then used to perform a weighted summation over frame representations, thereby integrating intra-modal semantic frame features:

$$c^{l+1}_{j,\text{intra}} = \sigma(\sum_{i \in \mathcal{N}_j} \beta^l_{ji,\text{intra}} W^l_{\text{intra}} h^{l+1,i}_{\text{intra}}). \quad (7)$$

Additionally, in our experiments, we use the multi-head attention mechanism to learn and integrate the intra-modal semantic frame features from multiple feature subspaces:

$$\begin{aligned} h^{l+1,i}_{\text{intra}} &= \Big\|_{k=1}^{K} \sigma(\sum_{j \in \mathcal{N}_i} \alpha^{l,k}_{ij,\text{intra}} W^{l,k}_{\text{intra}} c^l_j), \\ c^{l+1}_{j,\text{intra}} &= \Big\|_{k=1}^{K} \sigma(\sum_{i \in \mathcal{N}_j} \beta^{l,k}_{ji,\text{intra}} W^{l,k}_{\text{intra}} h^{l+1,i}_{\text{intra}}), \end{aligned} \quad (8)$$

where K denotes the number of attention heads.

Inter-modal FSA: Similarly, inter-modal FSA is designed to learn and integrate inter-modal semantic frames, thereby enabling a more nuanced capture of inter-modal sentiment cues. This process consists of two steps: Inter-modal SFL and Inter-modal SFI.

For inter-modal SFL, we adopt the same approach as for intra-modal SFL to update inter-modal semantic frame features:

$$h^{l+1,i}_{\text{inter}} = \Big\|_{k=1}^{K} \sigma(\sum_{j \in \mathcal{N}_i} \alpha^{l,k}_{ij,\text{inter}} W^{l,k}_{\text{inter}} c^l_j), \quad (9)$$

where $h^{l+1,i}_{\text{inter}}$ represents the updated inter-modal semantic frame features, $\alpha^{l,k}_{ij,\text{inter}}$ denotes the attention coefficient of the k-th head, and $W^{l,k}_{\text{inter}}$ is the learnable weight matrix for the k-th head.

For inter-modal SFI, we adopt the same approach to integrate inter-modal semantic frame features:

$$c^{l+1}_{j,\text{inter}} = \Big\|_{k=1}^{K} \sigma(\sum_{i \in \mathcal{N}_j} \beta^{l,k}_{ji,\text{inter}} W^{l,k}_{\text{inter}} h^{l+1,i}_{\text{inter}}), \quad (10)$$

where $c^{l+1}_{j,\text{inter}}$ represents the integrated inter-modal lexical unit features, and $\beta^{l,k}_{ji,\text{inter}}$ denotes the attention coefficient of the k-th head.

Fusion module: To comprehensively model intra- and inter-modal sentiment cues, we design a fusion mechanism to dynamically integrate the two types of information guided by semantic frames. The detailed approach is as follows.

$$\begin{aligned} p^l_{i,\text{intra}} &= \tanh(W^l_1 c^l_{j,\text{intra}} + b^l_1), \\ q^l_{i,\text{inter}} &= \tanh(W^l_2 c^l_{j,\text{inter}} + b^l_2), \\ \phi^l_i &= \sigma(W^l_3 c^l_{j,\text{intra}} + W^l_4 c^l_{j,\text{inter}} + b^l_3), \\ c^{l+1}_j &= \phi^l_i \odot p^l_{i,\text{intra}} + (1 - \phi^l_i) \odot q^l_{i,\text{inter}}, \end{aligned} \quad (11)$$

where $p^{l+1}_{i,\text{intra}}$ and $q^{l+1}_{i,\text{inter}}$ are transformed representations, $\phi^{(l+1)}_i$ is the gating vector, and $\odot$ denotes the Hadamard product.

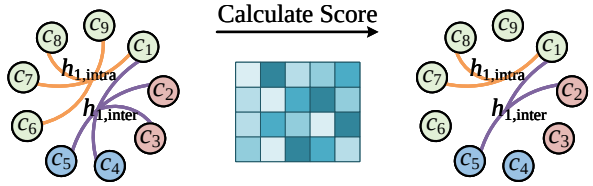

Fig. 2. Diagram of Dynamic Aware module

$W^l_*, (* \in \{1,2,3,4\})$ represents learnable weight matrices, while $b^l_*, (* \in \{1,2,3\})$ represents bias terms.

Dynamic aware module: During frame structure initialization, all frames are initially associated with all lexical units. However, a frame should be evoked only by specific lexical units that are semantically relevant rather than by arbitrary words. Therefore, we propose a dynamic-aware (DA) module, which evaluates the relevance between lexical units and frames, dynamically selects the most relevant units, and thereby reduces noise while enhancing representational capacity.

As shown in Fig. 2, the module computes an importance score function $s(i,j)$ for each frame to measure the strength of association between the i-th frame and the j-th lexical unit:

$$s(i,j) = \frac{\exp((W_f f_i)(W_x x_j)^T)}{\sum_{k \in \mathcal{N}_i} \exp((W_f f_i)(W_x x_k)^T)}, \quad (12)$$

where W_f and W_x are learnable weight matrices, f_i denotes the feature vector of the i-th frame, x_j refers to the feature vector of the j-th lexical unit in frame i, and N_i denotes the set of lexical units associated with frame i.

After computing the relevance scores $s(i,j)$, the dynamic-aware module adaptively filters lexical units according to the ranking of their scores. For each semantic frame, only the top-K proportion of lexical units with the highest relevance is retained, where K is a ratio hyperparameter controlling the strength of frame–lexical associations. This adaptive selection allows the model to dynamically preserve semantically salient units while suppressing irrelevant activations.

D. Semantic Fusion Module

The semantic fusion module implements the process of frame-based interpretation by integrating information from lexical units and semantic frames, and constructs a unified multimodal joint representation, thereby enhancing the model's capability for sentiment understanding.

However, the dynamic-aware process may introduce redundant frames (*i.e.*, multiple semantic frames associated with the same lexical unit), which not only increases computational costs but also leads to information redundancy. Therefore, we first remove the redundant frames before fusion.

To ensure the uniqueness of the semantic frames, we eliminate all redundant semantic frames, retaining only the first occurrence of each redundant semantic frame. The representation of the retained semantic frame is then updated by averaging the features of all redundant semantic frames, effectively fusing key information. Let the intra-modal semantic frame set and

inter-modal frame semantic set after deduplication be $E^o_{\text{intra}} \in \mathbb{R}^{N_{\text{intra}} \times T_m}$ and $E^o_{\text{inter}} \in \mathbb{R}^{N_{\text{inter}} \times T_m}$, respectively, where N_{intra} and N_{inter} denote the numbers of deduplicated intra-modal and inter-modal semantic frames, with their corresponding semantic frame representations denoted as:

$$\begin{aligned} H^o_{\text{intra}} &= \{h^{o,1}_{\text{intra}}, h^{o,2}_{\text{intra}}, \ldots, h^{o,N_{\text{intra}}}_{\text{intra}}\} \in \mathbb{R}^{N_{\text{intra}} \times d}, \\ H^o_{\text{inter}} &= \{h^{o,1}_{\text{inter}}, h^{o,2}_{\text{inter}}, \ldots, h^{o,N_{\text{inter}}}_{\text{inter}}\} \in \mathbb{R}^{N_{\text{inter}} \times d}. \end{aligned} \tag{13}$$

After semantic frame deduplication, we further update the semantic frame representations and lexical unit representations following the Intra-modal FSA and Inter-modal FSA methods described in Section III-C2, computed as follows:

$$\begin{aligned} h^{u,i}_{\text{intra}} &= \sum_{j \in \mathcal{N}_i} \alpha^o_{ij,\text{intra}} W^o_{\text{intra}} c^L_j, \\ h^{u,i}_{\text{inter}} &= \sum_{j \in \mathcal{N}_i} \alpha^o_{ij,\text{inter}} W^o_{\text{inter}} c^L_j, \\ c^{u,j}_{\text{intra}} &= \sum_{i \in \mathcal{N}_j} \beta^o_{ji,\text{intra}} W^o_{\text{intra}} h^{o,i}_{\text{intra}}, \\ c^{u,j}_{\text{inter}} &= \sum_{i \in \mathcal{N}_j} \beta^o_{ji,\text{inter}} W^o_{\text{inter}} h^{o,i}_{\text{inter}}, \end{aligned} \tag{14}$$

where $h^{u,i}_{\text{intra}}$ and $h^{u,i}_{\text{inter}}$ represent the updated intra-modal and inter-modal frame representations, $c^{u,j}_{\text{intra}}$ and $c^{u,j}_{\text{inter}}$ represent the updated intra-modal and inter-modal lexical unit representations, $\beta^o_{ji,\text{intra}}$ and $\beta^o_{ji,\text{inter}}$ denote the attention weights for intra-modal and inter-modal interactions, respectively. W^o_{intra} and W^o_{inter} are learnable weight matrices, N_i denotes the set of lexical units associated with the target frame, while N_j also refers to the set of frames evoked by a given lexical unit.

We use $H^u_{\text{intra}} \in \mathbb{R}^{N_{\text{intra}} \times d}$ and $H^u_{\text{inter}} \in \mathbb{R}^{N_{\text{inter}} \times d}$ to represent the updated intra- and inter-modal semantic frame features, respectively. Similarly, $C^u_{\text{intra}} \in \mathbb{R}^{N_T \times d}$ and $C^u_{\text{inter}} \in \mathbb{R}^{N_T \times d}$ denote the updated intra- and inter-modal lexical unit representations, respectively. Finally, we adopt mean pooling to these features and concatenate them to obtain the multimodal joint representation.

$$o = \left[\mathrm{P}\left(H^u_{\text{intra}}\right) || \mathrm{P}\left(H^u_{\text{inter}}\right) || \mathrm{P}\left(C^u_{\text{intra}}\right) || \mathrm{P}\left(C^u_{\text{inter}}\right)\right], \tag{15}$$

where $o \in \mathbb{R}^{4d}$ represents the multimodal joint representation, and $P(\cdot)$ denotes mean pooling.

E. Predict and Train

We use two fully connected layers to predict the sentiment labels and compute the loss using the L_1 loss function during training. The process can be expressed as follows. Let the predicted label be $\hat{y}$ and the ground truth label be y. The prediction is computed as:

$$\hat{y} = W^o_2 \left(\tanh\left(W^o_1 o + b^o_1\right)\right) + b^o_2, \tag{16}$$

where W^o_1 and W^o_2 represent the weight matrices for the first and second fully connected layers, b^o_1 and b^o_2 are the corresponding bias terms, and $\tanh(\cdot)$ is the activation function. These two layers help the model to map the input features to the predicted sentiment labels. Accordingly, the loss is computed as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{17}$$

where N is the mini-batch size, $\hat{y}_i$ is the predicted label, and y_i is the ground truth label. The L1 loss is used to minimize the absolute differences between the predicted and true labels. This loss function is used for model training.

TABLE II
BASIC STATISTICS OF THE DATASETS.

Dataset	Train	Valid	Test	Total
CMU-MOSI	1284	229	686	2119
CMU-MOSEI	16326	1871	4659	23453
CH-SIMS	1368	456	457	2281

IV. EXPERIMENT

In this section, we conduct extensive experiments to validate the effectiveness of the proposed model. Besides, ablation studies, sensitivity tests, and case studies have also been conducted to further reveal the adaptability of the proposed approach.

A. Datasets

In this paper, we use the CMU-MOSI [34], CMU-MOSEI [26], and CH-SIMS [35] datasets to evaluate the proposed model. Table II summarizes the basic statistics of the datasets.

CMU-MOSI [34] is a widely used benchmark dataset in the field of multimodal sentiment analysis, focusing on opinion-based sentiment recognition tasks. The dataset is collected from YouTube and contains 93 review videos recorded by 89 reviewers, with a total of 2,199 independent video segments. Each segment includes information from three modalities: text, visual, and audio, and is annotated with sentiment intensity scores, ranging from -3 (strongly negative) to $+3$ (strongly positive).

CMU-MOSEI [26] is an extended version of the CMU-MOSI dataset for multimodal sentiment analysis, offering a larger and more diverse set of multimodal data. It contains 23,453 video segments from 3,228 videos covering 250 different topics, recorded by 1,000 distinct YouTube speakers. Each video segment is labeled with sentiment intensity scores (ranging from -3 to $+3$) and is further annotated with emotional categories based on Ekman's six basic emotions (happiness, sadness, anger, fear, disgust, and surprise).

CH-SIMS [35] is a publicly available Chinese multimodal sentiment analysis dataset specifically designed for opinion-based sentiment analysis tasks. The dataset contains 2,281 video segments from 60 original videos, sourced from movies, TV series, and variety shows, focusing on emotional expression and opinion-based sentiment analysis. Each video segment is manually annotated with sentiment scores, ranging from -1 (strongly negative) to $+1$ (strongly positive). To support multimodal sentiment analysis, the dataset includes

text, audio, and visual information, making it suitable for emotion recognition and analysis.

The selection of these datasets is based on their representativeness in terms of task types, multimodal characteristics, and linguistic coverage. Together, they span core subtasks in multimodal sentiment analysis and rang from opinion-based sentiment classification to sentiment intensity estimation, and cover both English and Chinese languages, providing a comprehensive benchmark for evaluating model generalizability in diverse real-world scenarios.

To comprehensively evaluate the performance of the proposed model on the above datasets, we adopt a set of widely used evaluation metrics, and we follow the task settings established in previous studies [35], [36]. For the CMU-MOSI and CMU-MOSEI datasets, we use 7-class accuracy (Acc-7), binary classification accuracy (Acc-2), F1-score, Mean Absolute Error (MAE), and Pearson Correlation (Corr). Specifically, Acc-2 and F1 are computed under two binary classification schemes, denoted as Negative vs. non-negative/Negative vs. positive in the experimental results, to evaluate sentiment polarity discrimination. MAE and Corr are used to assess the model's performance in sentiment intensity regression. For the CH-SIMS dataset, we use binary classification accuracy (Acc-2), 3-class accuracy (Acc-3), 5-class accuracy (Acc-5), and F1-score, along with MAE and Corr for evaluating sentiment intensity prediction.

It is important to note that for the metrics mentioned above, a lower MAE indicates smaller prediction errors and better performance, while higher values for the other metrics indicate better model performance.

B. Experimental Setup

1) Feature Extraction: In this experiment, we used BERT [37], COVAREP [38], Facet[1], LibROSA [39], and OpenFace [40] as feature extractors. For CMU-MOSI and CMU-MOSEI, text features were extracted using the pre-trained BERT model, resulting in 768-dimensional vectors. Audio features were extracted using COVAREP, yielding 74-dimensional features, which include Mel-Frequency Cepstral Coefficients (MFCCs), glottal source parameters, and peak slope, among others. Visual features were extracted using the Facet tool, providing 35-dimensional features, including facial landmark points, seven basic emotions, and action units. For CH-SIMS, text features were similarly extracted using the pre-trained BERT model with a dimensionality of 768, consistent with CMU-MOSI and CMU-MOSEI. Audio features were extracted using the LibROSA speech toolkit, providing 33-dimensional frame-level features, including 20 MFCCs, 12 CQT (Constant-Q Transform) coefficients, and 1 log fundamental frequency (log F0). Visual features were obtained through MTCNN [41] for face detection, and the MultiComp OpenFace 2.0 toolkit was used to extract 709-dimensional features, covering facial landmark points, action units, head poses, and eye gaze information.

Following the benchmark protocol of the MMSA framework, we use the standard processed features and data splits for CMU-MOSI, CMU-MOSEI, and CH-SIMS to ensure fair comparison. CMU-MOSI and CMU-MOSEI are evaluated with aligned features, while CH-SIMS is evaluated with unaligned features. For DAFSA, the supervision signal is the sample-level multimodal sentiment label, and no manually defined modality-reliability prior is introduced.

[1]https://imotions.com/

TABLE III
HYPERPARAMETER SETTINGS FOR THE MODEL ACROSS CMU-MOSI, CMU-MOSEI, AND CH-SIMS DATASETS.

Perparameter	CMU-MOSI	CMU-MOSEI	CH-SIMS
Max epoch	50	50	50
Batch size	32	32	16
Learning rate	1e-5	1e-5	8e-6
Early stopping	8	8	8
Model dim	512	512	512
L	2	4	2
K	0.4	0.6	0.5
Heads	4	8	4

2) Baseline: We compare DAFSA with representative MSA baselines from three categories. The non-graph-based methods include TFN [18] and LMF [19] for tensor-based or low-rank fusion, MulT [32] and CENet [42] for attention-based interaction, MISA [36] for representation disentanglement, and Self-MM [24], MMML [43], and CLGSI [44] for self-supervised, multi-loss, or contrastive learning. The graph-based methods include Graph-MFN [26], MTAG [31], and SSLMM-CM [30], which model multimodal dependencies with graph structures. The hypergraph-based methods include MDH [45] and MHN-CL [46], which capture high-order multimodal relations. Detailed descriptions are provided in Appendix A.

3) Environment Configuration: For fair comparison, all experiments are conducted under the standard benchmark protocol and unified evaluation settings. We train DAFSA using the Adam optimizer [47] with a learning-rate decay strategy, and the key hyperparameter settings for each dataset are summarized in Table III. Detailed information about baseline result sources, additional implementation settings, and hardware/software environments is provided in Appendix B.

C. Comparative Experiments

To thoroughly evaluate the effectiveness of our proposed model, we conduct two sets of comparative experiments. The first compares our method with widely adopted multimodal sentiment analysis baselines, while the second investigates its performance against representative large language models (LLMs) to assess its competitiveness in the era of foundation models.

1) Comparison with MSA Baselines: All MSA baseline experiments are conducted under five random seeds, i.e., [1111, 1112, 1113, 1114, 1115]. Table IV and Table V report the mean results, and the complete mean $\pm$ standard deviation results are provided in Appendix C.

From Table IV and Table V, it can be observed that DAFSA achieves the best or second-best across multiple metrics on all three datasets. Specifically, DAFSA reaches the highest performance on the CMU-MOSI dataset in terms of Acc-2,

TABLE IV
PERFORMANCE COMPARISON OF THE DAFSA AND BASELINE ON THE CMU-MOSI AND CMU-MOSEI (**BOLD** INDICATES THE BEST, AND UNDERLINE INDICATES THE SECOND-BEST).

Models	CMU-MOSI					CMU-MOSEI				
	Acc-2	F1-score	Acc-7	MAE	Corr	Acc-2	F1-score	Acc-7	MAE	Corr
TFN	76.88/77.93	76.78/77.90	34.49	0.968	0.657	78.71/81.71	79.12/81.54	51.67	0.573	0.719
LMF	78.22/79.48	78.11/79.43	35.60	0.922	0.677	79.14/83.47	79.80/83.49	52.16	0.567	0.734
MulT	79.10/80.46	79.08/80.50	33.67	0.934	0.688	81.05/84.24	81.47/84.15	52.73	0.560	0.733
MISA	81.98/83.72	81.93/83.73	42.30	0.770	0.776	80.90/84.66	81.42/84.64	52.72	0.550	0.758
Self-MM	82.10/83.69	82.04/83.68	46.15	**0.732**	0.786	82.22/84.79	82.54/84.68	53.19	**0.535**	0.766
MMML	82.80/84.91	82.71/84.89	43.73	0.735	0.785	82.70/84.78	82.90/84.60	51.84	0.553	**0.768**
CENet	82.13/83.84	82.03/83.80	43.99	0.743	0.783	81.74/85.65	82.21/85.60	53.73	0.539	0.733
CLGSI	82.22/84.60	82.05/84.53	43.44	0.755	0.779	83.24/85.11	83.39/84.91	53.66	0.537	0.763
Graph-MFN	76.79/78.02	76.71/78.01	34.40	0.966	0.653	80.27/83.76	80.77/83.70	52.00	0.564	0.727
MTAG	80.76/83.08	80.39/82.82	25.86	0.879	0.706	81.52/84.64	81.70/84.53	52.59	0.559	0.735
SSLMM-CM	83.38/85.09	83.45/85.37	46.51	0.733	**0.798**	83.64/85.93	83.70/86.05	53.20	0.551	0.759
MDH	81.02/82.47	81.10/82.52	38.39	0.893	0.694	82.52/84.49	82.65/84.25	50.41	0.556	0.733
MHN-CL	82.21/84.67	82.32/83.76	39.45	0.801	0.726	83.31/85.95	83.40/86.02	51.42	0.549	0.745
DAFSA	**83.94/86.28**	**83.71/86.14**	**46.75**	**0.732**	0.782	**84.05/86.36**	**84.09/86.38**	**53.85**	0.541	**0.768**

TABLE V
PERFORMANCE COMPARISON OF THE DAFSA AND BASELINE ON THE CH-SIMS DATASET (**BOLD** INDICATES THE BEST, AND UNDERLINE INDICATES THE SECOND-BEST).

Model	Acc-2	Acc-3	Acc-5	F1-score	MAE	Corr
TFN	76.89	64.99	39.56	77.22	0.437	0.577
LMF	77.38	65.38	39.91	77.25	0.443	0.567
MulT	78.16	65.78	39.74	78.10	0.438	0.585
MISA	76.67	64.07	38.38	76.81	0.445	0.566
Self-MM	78.34	64.82	40.48	78.26	0.426	0.584
MMML	77.46	65.21	**44.86**	77.76	0.431	0.565
CENet	77.52	61.58	33.25	77.14	0.469	0.540
CLGSI	78.77	66.40	44.42	78.62	0.427	0.604
Graph-MFN	76.93	64.81	40.70	76.92	0.438	0.570
MTAG	77.02	54.27	34.14	75.59	0.478	0.517
SSLMM-CM	80.06	66.08	40.87	80.16	0.429	0.596
MDH	79.21	65.43	41.04	79.27	0.422	0.608
MHN-CL	81.05	65.68	41.21	**81.02**	0.419	0.610
DAFSA	**81.40**	**66.96**	**44.86**	80.99	**0.417**	**0.619**

F1-score, Acc-7, and MAE, with improvements of 0.56/1.19, 0.26/0.77, 0.24, and 0.001 over the second-best, respectively. On the CMU-MOSEI dataset, it performs optimally on Acc-2, F1-score, Acc-7, and Corr, exceeding the next-best model by 0.41/0.41, 0.39/0.33, 0.12, and 0.002, respectively. On the CH-SIMS dataset, DAFSA achieves the best results on Acc-2, Acc-3, Acc-5, MAE, and Corr, with respective gains of 0.35, 0.56, 0.44, 0.002, and 0.009 over the runner-up.

In comparison with baseline methods. Non-graph-based methods (TFN, LMF, MulT, MISA, Self-MM, MMML, CENet and CLGSI) perform well on standard sentiment analysis tasks but are limited in capturing subtle emotional variations. They primarily focus on extracting explicit emotional features, often missing the nuanced changes in sentiment and the deeper interrelations between emotions.

Graph-based methods (Graph-MFN, SSLMM-CM and MTAG) leverage graph structures to model inter-modality dependencies. While they can effectively capture basic interactions, their ability to model higher-order emotional interactions and subtle sentiment changes remains constrained, as graphs typically only capture second-order relationships.

Hypergraph-based methods (MDH, MHN-CL), on the other hand, can model more complex high-order dependencies by introducing hyperedges, enabling them to represent intricate multimodal interactions. However, they tend to oversimplify subtle emotional fluctuations and tonal changes, limiting their effectiveness in modeling finer emotional details.

In summary, we propose the DAFSA method, which introduces the concept of semantic frames. By exploring and integrating semantic frames, DAFSA is able to more accurately capture the subtle differences in emotional information, thereby effectively enhancing the performance and interpretability of multimodal sentiment analysis.

2) Comparison with Large Language Models: In this section, we select four open-source text-based LLMs: Qwen2.5 [48], Baichuan2 [49], LLaMA3.1[2], and ChatGLM3[3], which are accessed via their publicly available API interfaces without any additional fine-tuning. Since these models are not specifically designed for multimodal sentiment analysis tasks, we only provide the textual modality of each input sample and adopt a prompt-based inference strategy for sentiment prediction. Each model receives a predefined prompt such as: "Please analyze the following sentence and choose a number from $-x$ to $+x$ to represent its sentiment intensity: {text}". Here, we set $x = 3$ for CMU-MOSI and CMU-MOSEI, and $x = 1$ for CH-SIMS, in alignment with their respective label distributions. These general-purpose LLMs are chosen because, to the best of our knowledge, there currently exist no publicly available, API-accessible multimodal large models capable of simultaneously processing text, audio, and video inputs.

The experimental results are presented in Table VI and Table VII, which respectively report the evaluation results of the models on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets.

[2]https://www.llama.com/models/llama-3
[3]https://github.com/THUDM/ChatGLM-6B

TABLE VI
COMPARISON OF LLMS RESULTS ON THE CMU-MOSI AND CMU-MOSEI (**BOLD** INDICATES THE BEST, AND UNDERLINE INDICATES THE SECOND-BEST).

Models(Size)	CMU-MOSI					CMU-MOSEI				
	Acc-2	F1-score	Acc-7	MAE	Corr	Acc-2	F1-score	Acc-7	MAE	Corr
Qwen2.5(14B)	79.88/85.32	80.04/85.24	40.09	0.920	0.772	**84.50**/85.51	**84.37**/84.57	40.09	0.920	0.739
BaiChuan2(13B)	61.95/64.95	70.08/76.46	22.30	1.321	0.370	44.49/58.76	47.38/67.55	19.78	1.224	0.350
Llama3.1(8B)	80.00/82.08	80.58/84.57	13.97	2.287	0.611	79.82/84.38	79.11/85.42	27.30	1.891	0.546
ChatGLM3(6B)	77.55/78.86	77.54/78.64	34.40	1.580	0.203	81.29/70.20	81.31/71.26	30.12	0.974	0.482
DAFSA(0.118B/0.121B)	**83.94/86.28**	**83.71/86.14**	**46.75**	**0.732**	**0.782**	84.05/**86.36**	84.09/**86.38**	**53.85**	**0.541**	**0.768**

TABLE VII
COMPARISON OF LLMS RESULTS ON THE CH-SIMS (**BOLD** INDICATES THE BEST, AND UNDERLINE INDICATES THE SECOND-BEST).

Model(Size)	Acc-2	Acc-3	Acc-5	F1-score	MAE	Corr
Qwen2.5(14B)	78.12	56.67	37.42	80.79	0.440	**0.634**
BaiChuan2(13B)	69.58	43.54	20.57	78.42	0.545	0.325
Llama3.1(8B)	72.21	51.64	33.92	**81.13**	0.556	0.428
ChatGLM3(6B)	69.15	43.53	29.98	73.16	0.531	0.394
DAFSA(0.1128)	**81.40**	**66.96**	**44.86**	80.99	**0.417**	0.619

From Table VI and Table VII, we observe that our proposed model consistently achieves either the best or second-best performance across most evaluation metrics on all three datasets, demonstrating its effectiveness in sentiment analysis tasks. Specifically, on the CMU-MOSI dataset, our model achieves the best performance across all evaluation metrics, outperforming the second-best model by 3.94/0.96 in Acc-2, 3.13/0.90 in F1-score, 6.66 in Acc-7, 0.188 in MAE, and 0.01 in Corr, respectively. On CMU-MOSEI, our method ranks first in Acc-2 (Negative vs. Positive), F1-score (Negative vs. Positive), Acc-7, MAE, and Corr, outperforming the runner-up by 0.85, 1.81, 13.76, 0.379, and 0.029, respectively. On the CH-SIMS dataset, our model achieves top results on Acc-2, Acc-3, Acc-5, and MAE, with improvements of 3.28 in Acc-2, 10.29 in Acc-3, 7.44 in Acc-5, and 0.023 in MAE over the next-best model. These results suggest that DAFSA maintains competitive performance across different languages, annotation granularities, and task configurations.

In addition, mainstream large language models typically contain tens of billions of parameters. In contrast, our proposed model, DAFSA, has an average parameter size of only 0.117B (117 million), making it approximately 100× smaller in scale. Despite this substantial disparity in model size, DAFSA achieves comparable or superior performance across all three sentiment analysis datasets, highlighting the effectiveness of its task-specific architecture and sentiment reasoning mechanism. These findings indicate that DAFSA is capable of providing efficient and reliable sentiment understanding without relying on large-scale computational resources, demonstrating strong potential for practical deployment and real-world applications.

D. Ablation Study

In order to verify the source of DAFSA's performance gains, we conduct two sets of ablation studies in this section. The module ablation study examines the necessity of the key components in semantic frame construction, learning, integration, and dynamic updating, while the modality ablation study evaluates the contribution of different modalities to frame-semantic sentiment reasoning.

1) Module Ablation Study: To assess the effectiveness of the key techniques introduced in the mathematical modeling of frame semantics, we design a module ablation study, with the specific design outlined as follows.

***w/o* gating:** Remove the gating fusion module and use the average of intra-modal and inter-modal lexical unit features instead. This setting is used to verify whether adaptive fusion is necessary for balancing intra-modal and inter-modal frame information. ***w/o* attention:** Remove the attention mechanism in the SFL and SFI stages and update and integrate semantic frames using weighted summation. This setting examines whether relevance-aware interaction between lexical units and semantic frames is necessary for effective frame learning and integration. ***w/o* DA:** Remove the DA module and retain redundant lexical unit features in semantic frames. This setting evaluates whether dynamic filtering of irrelevant or redundant lexical units contributes to more effective frame structure refinement. ***w/o* FS:** Remove the FS module and concatenate features after applying global average pooling to both lexical unit features and semantic frame features. This setting is used to verify whether the proposed semantic fusion process is necessary for integrating lexical-level and frame-level information. ***w/o* SFL:** Remove the SFL stage, halting the update process of semantic frames and leaving only the semantic frame integration stage. This setting examines whether updating semantic frames from their associated lexical units is necessary for contextual frame refinement. ***w/o* SFI:** Remove the SFI stage, halting the integration process of semantic frames and leaving only the semantic frame learning stage. This setting evaluates whether propagating learned frame information back to lexical unit representations is necessary for sentiment reasoning. ***w/o* SFL & SFI:** Remove both the semantic frame learning and semantic frame integration stages, using only lexical unit features for sentiment prediction. This setting is designed to evaluate whether the bidirectional interaction between lexical units and semantic frames is essential for frame-semantic sentiment reasoning. ***w/o* FS & DA:** Remove both the FS and DA modules, without dynamic frame refinement or the proposed semantic fusion process. This setting is used to examine the complementarity between frame structure refinement and lexical-frame semantic fusion.

TABLE VIII
MODULE ABLATION STUDY RESULTS ON THE CMU-MOSI AND CMU-MOSEI.

Models	CMU-MOSI					CMU-MOSEI				
	Acc-2	F1-score	Acc-7	MAE	Corr	Acc-2	F1-score	Acc-7	MAE	Corr
w/o gating	82.92/84.76	82.94/84.82	44.23	0.758	0.774	82.71/84.89	82.56/85.40	53.20	0.564	0.758
w/o attention	82.90/84.45	82.67/84.33	43.31	0.753	0.770	82.65/84.85	82.52/85.33	53.11	0.560	0.760
w/o DA	83.20/85.19	83.11/85.15	45.66	0.748	0.768	82.78/85.62	82.30/85.93	53.26	0.545	0.762
w/o FS	82.63/84.60	82.56/84.64	42.46	0.754	0.775	83.01/85.82	83.10/85.76	53.05	0.554	0.761
w/o SFL	82.62/84.25	82.48/84.33	41.77	0.771	0.769	82.34/85.58	82.45/85.48	52.58	0.574	0.752
w/o SFI	82.57/84.30	82.35/84.41	40.85	0.760	0.772	82.41/85.99	82.81/85.94	52.31	0.565	0.759
w/o SFI & SFL	76.46/78.30	76.16/78.11	37.40	0.888	0.652	80.35/83.70	80.22/81.63	50.64	0.599	0.729
w/o FS & DA	82.24/84.24	82.32/84.20	41.88	0.764	0.738	82.45/84.82	82.07/84.68	52.70	0.571	0.754
DAFSA	**83.94/86.28**	**83.71/86.14**	**46.75**	**0.732**	**0.782**	**84.05/86.36**	**84.09/86.38**	**53.85**	**0.541**	**0.768**

TABLE IX
MODULE ABLATION STUDY RESULTS ON THE CH-SIMS.

Model	Acc-2	Acc-3	Acc-5	F1-score	MAE	Corr
w/o gating	80.35	65.55	43.92	79.85	0.435	0.610
w/o attention	80.46	65.96	43.27	79.52	0.432	0.604
w/o DA	79.59	65.21	42.74	79.35	0.440	0.585
w/o FS	79.87	65.74	43.98	78.51	0.436	0.597
w/o SFL	78.12	63.68	40.48	77.30	0.459	0.574
w/o SFI	78.99	64.33	42.01	78.77	0.439	0.575
w/o SFI & SFL	76.35	61.74	38.56	75.58	0.482	0.542
w/o FS & DA	79.08	64.87	42.21	78.08	0.453	0.576
DAFSA	**81.40**	**66.96**	**44.86**	**80.99**	**0.417**	**0.619**

The experimental results, as shown in Table VIII and Table IX, indicate that removing any module leads to a decrease in performance, further confirming the importance of each module in the process of information modeling and representation. Specifically, the gating module is used to dynamically fuse intra-modality and inter-modality lexical unit features. Its removal limits the fusion effectiveness of the lexical unit features. The attention mechanism in the SFL and SFI stages adjusts the relevance between the semantic frame and lexical units, highlighting key information; without it, the model struggles to accurately capture sentiment-related semantic units. The DA module removes redundant lexical unit features during the construction of the semantic frame, and its removal introduces ineffective features, reducing the model's performance. The SFL stage updates the semantic frame to adapt to the input context. Its removal weakens the frame's responsiveness to the input. The SFI stage integrates the semantic frame into the lexical unit features, and its absence impacts the comprehensiveness of sentiment information. Finally, the FS module integrates the semantic frame with lexical unit features into a unified multimodal representation. Its removal directly affects the model's ability to express multimodal semantics, leading to a significant drop in performance.

Beyond single-component ablations, the combined ablation results further reveal the interactions among related modules. Compared with removing SFL or SFI individually, *w/o* SFL & SFI causes a more evident performance drop, suggesting that semantic frame learning and semantic frame integration are complementary processes. SFL updates semantic frames from lexical units, whereas SFI integrates frame-level information back into lexical representations; removing both weakens the bidirectional interaction between lexical units and semantic frames. Similarly, *w/o* FS & DA leads to a larger performance decrease than removing either FS or DA alone, indicating that dynamic frame refinement and semantic fusion jointly contribute to effective frame-semantic sentiment reasoning. These results demonstrate that the proposed modules are not independent architectural additions. Instead, they correspond to different stages of semantic frame construction, updating, refinement, and integration. Their individual contributions and mutual interactions jointly support the overall performance of DAFSA.

2) Modality Ablation Study: To further investigate the role of each input modality in MSA, we design a modality ablation study by selectively removing one or more modalities during the training and testing stages. The specific experimental settings are detailed as follows.

Single-modality: Using only one modality, namely text (T), audio (A) or visual (V). **Dual-modality combination:** Using a combination of two modalities (T+A, T+V or A+V). **Tri-modality fusion:** Using all three modalities simultaneously (T+A+V). In both single- and dual-modality experiments, the unused modalities are directly removed to avoid introducing additional noise. All experiments adopt the same model framework, with adjustments made only to the input branches according to the number of modalities used. The experimental results are presented in Table X and Table XI.

Based on Table X and Table XI, we draw the following key findings. For single-modality performance, across all datasets, the text modality consistently achieved the best performance, indicating that textual information plays a dominant role in sentiment analysis, which is also consistent with the conclusions of previous studies [50], [51]. This phenomenon suggests that the text modality can directly activate the core elements of sentiment frames (such as sentiment categories and polarity), thereby providing a relatively complete and stable semantic basis even under single-modality conditions. For dual-modality combination performance: in all three datasets, dual-modality combinations involving the text modality (T+A, T+V) consistently outperformed the pure non-text combination (A+V). This is because effective sentiment understanding relies on

TABLE X
MODALITY ABLATION STUDY RESULTS ON THE CMU-MOSI AND CMU-MOSEI.

Models	CMU-MOSI					CMU-MOSEI				
	Acc-2	F1-score	Acc-7	MAE	Corr	Acc-2	F1-score	Acc-7	MAE	Corr
T	83.05/85.51	82.90/85.44	42.75	0.799	0.774	81.84/84.85	82.12/84.97	52.35	0.564	0.757
A	58.23/57.98	57.72/57.65	17.91	1.343	0.235	69.19/67.26	65.60/64.14	42.49	0.803	0.189
V	52.16/50.87	49.55/49.36	17.29	1.417	0.098	68.59/66.38	67.05/64.92	42.74	0.793	0.249
T+A	83.27/85.76	83.11/85.65	44.68	0.736	0.779	83.32/85.85	83.03/85.80	52.81	0.555	0.760
T+V	83.20/85.61	82.97/85.56	44.45	0.748	0.777	82.74/85.71	83.24/85.90	51.82	0.551	0.764
A+V	59.70/59.45	58.12/57.90	18.23	1.333	0.240	70.18/67.47	67.60/65.85	43.05	0.779	0.259
T+A+V	**83.94/86.28**	**83.71/86.14**	**46.75**	**0.732**	**0.782**	**84.05/86.36**	**84.09/86.38**	**53.85**	**0.541**	**0.768**

TABLE XI
MODALITY ABLATION STUDY RESULTS ON THE CH-SIMS.

Model	Acc-2	Acc-3	Acc-5	F1-score	MAE	Corr
T	77.35	64.65	39.39	79.17	0.436	0.569
A	65.44	47.59	21.44	64.87	0.547	0.159
V	71.29	52.75	22.54	59.25	0.559	0.015
T+A	78.03	66.22	42.64	80.48	0.426	0.587
T+V	78.71	66.35	43.45	79.71	0.433	0.575
A+V	72.82	54.26	24.04	66.40	0.519	0.178
T+A+V	**81.40**	**66.96**	**44.86**	**80.99**	**0.417**	**0.619**

the joint activation of conceptual cues (provided by text) and perceptual cues (provided by audio or visual modalities), which enriches the depth and accuracy of sentiment reasoning. For tri-modality fusion performance: Across all datasets, tri-modality fusion (T+A+V) achieved the best overall performance. This result suggests that complete and accurate sentiment understanding typically depends on the collaborative activation of multiple modalities. The text modality provides the basic semantic skeleton for the sentiment frame, the audio modality supplements variations in sentiment intensity and speaker attitude, and the visual modality primarily contributes non-verbal sentiment cues through facial expression features. The three modalities work together to complete the construction of various elements within the sentiment frame, significantly enhancing the overall performance of the sentiment analysis system.

In summary, the proposed frame-semantic-based modeling approach provides a structured way to analyze the collaboration of multimodal information and offers additional interpretability evidence for sentiment analysis.

E. Parameter Sensitivity

In this section, we conduct parameter sensitivity experiments to analyze the robustness of DAFSA with respect to the hyperparameter K. Here, K is a key parameter in the DA module, which controls the retention ratio of lexical units for each semantic frame and thereby affects the sparsity of lexical-frame associations.

To analyze the impact of the value of K in the DA module on model performance, we conducted experiments with $K = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. The experimental results, as shown in Fig. 3, reveal that as K increases, the model performance follows a 'first increasing, then decreasing' trend. Specifically, when K is small, the filtering is too strict, leading to the loss of emotional information; when K is too large, information redundancy increases, which interferes with the effective modeling of the semantic frame. However, within the middle range, the model achieves a good balance between retaining key information and suppressing redundancy, thereby achieving optimal performance. Moreover, in cross-dataset comparison experiments, the optimal K values on CMU-MOSI, CMU-MOSEI, and CH-SIMS are 0.4, 0.6, and 0.5, respectively. This phenomenon indicates that different datasets have differences in sample complexity, modality consistency, and linguistic structure, which affect the selection of the optimal token selection ratio. This further validates the adaptability and adjustability of the TopK strategy under different task conditions. Additional sensitivity analyses of the number of SFPU layers and attention heads are provided in Appendix E.

F. Visualization Analysis

The visualization analysis is conducted to further examine whether semantic frame modeling improves representation organization and provides interpretable frame-level evidence for sentiment reasoning.

1) Feature Visualization of Semantic Frame Integration: To further investigate the impact of SFI module in DAFSA on model performance, we conduct dimensionality reduction and visualization using t-SNE [52] on the multimodal joint representations extracted from the CMU-MOSI test set, as illustrated in Fig. 4. The visualization results reveal that the SFI module substantially improves the organization of the multimodal feature space. Specifically, SFI enhances intra-class compactness and inter-class separability by producing more coherent sentiment clusters with clearer boundaries. Neutral samples form a smooth transition region between positive and negative classes, reflecting the ability of semantic frames to encode sentiment continuity. Collectively, these effects demonstrate that SFI strengthens both the structural coherence and interpretability of the learned representations, enabling the model to capture fine-grained affective distinctions more effectively.

2) Semantic Frame Visualization: To compare the modeling capabilities of SSLMM-CM and DAFSA, and considering that text analysis is relatively easier, we randomly select a test sample from the CMU-MOSI dataset. We then visualize the correlation matrix of the text modality in SSLMM and

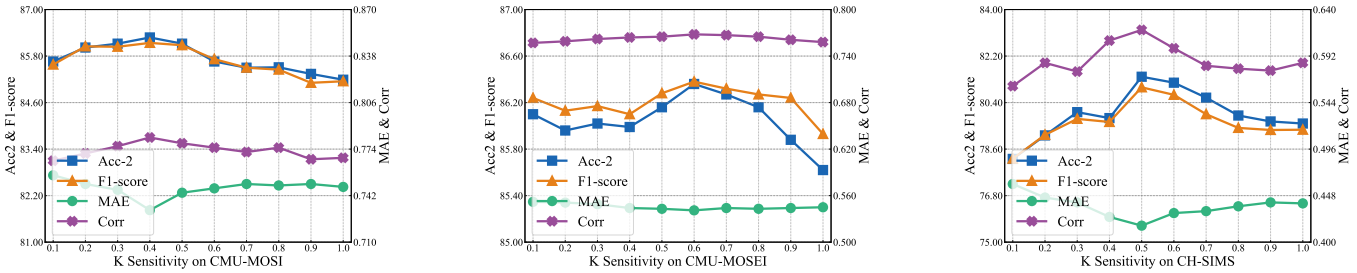

Fig. 3. Sensitivity analysis of K on different datasets.

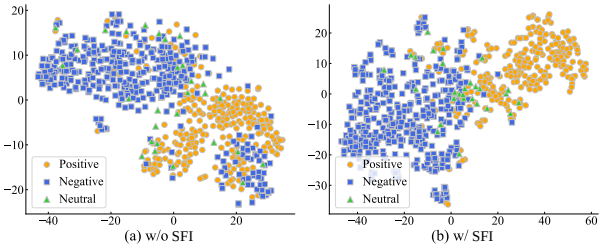

Fig. 4. Feature space visualization of CMU-MOSI test samples without SFI (a) and with SFI (b).

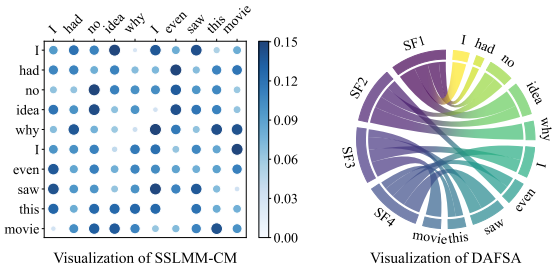

Fig. 5. Visualization of the text modality: Relevance distribution in SSLMM-CM (Left) and semantic frame representations in DAFSA (Right).

the semantic frame in DAFSA. The results are shown in Fig. 5. As shown in Fig. 5, although SSLMM can model the strength of correlations between words, it lacks the capacity for comprehensive analysis across multiple sentiment contexts. In contrast, DAFSA extracts four semantic frames (SF1, SF2, SF3, SF4) from the same sample and performs more structured sentiment analysis based on them. Specifically, SF1 includes the lexical units "I", "had", "no", and "idea", where the negated semantic construct (no idea) conveys a sense of helplessness and negative confusion. SF2, composed of "no", "idea", "why", and "even", reflects the frustration of failed reasoning, with the intensifier even reinforcing sentiments of self-doubt and sarcasm. SF3 consists of "I", "even", "saw", and "this", expressing regret and emotional volatility through a rhetorical structure of "unexpectedly doing something disappointing". Finally, SF4 includes "saw", "this", "movie", and "I", emphasizing movie as the main evaluative object and indicating a strong negative sentiment toward it. By integrating these frames, it can be inferred that the statement expresses strong negative sentiment.

G. Summary of Performance and Limitations

The proposed DAFSA model achieves competitive or superior results across CMU-MOSI, CMU-MOSEI, and CH-SIMS, demonstrating the effectiveness of mathematically modeling frame semantics for multimodal sentiment understanding. The ablation studies, parameter sensitivity analysis, and visualization results further show that the proposed frame-semantic modeling mechanism contributes to stable and interpretable sentiment reasoning. Additional Chinese semantic frame visualization and qualitative error analysis are provided in Appendix F and Appendix G, respectively.

Although DAFSA achieves strong performance, it still introduces additional computational overhead compared with lightweight baselines because of multi-level attention and frame-updating operations. A detailed computational cost analysis is provided in Appendix D. In addition, the Dynamic-Aware module retains lexical units according to a fixed proportion, which may limit its flexibility in adapting to diverse contextual structures. Future work will explore adaptive frame allocation and lightweight optimization to improve efficiency while preserving interpretability.

V. Conclusion

In this paper, we introduced frame semantics into multimodal sentiment analysis and proposed Dynamic-Aware Frame Semantic Analysis (DAFSA). DAFSA represents multimodal sequential features as lexical units and organizes them through intra-modal and inter-modal semantic frames, enabling sentiment reasoning from the perspective of frame-semantic organization. Through frame structure initialization, frame semantic learning, dynamic-aware updating, and semantic fusion, DAFSA models the evocation, refinement, and integration of multiple semantic frames in multimodal contexts. Experimental results on CMU-MOSI, CMU-MOSEI, and CH-SIMS show that DAFSA achieves competitive or improved performance across multiple evaluation metrics, providing empirical support for the proposed frame-semantic modeling framework. Further ablation studies, sensitivity analysis, computational cost analysis, and visualization analyses provide additional evidence for the effectiveness, robustness, efficiency, and interpretability of DAFSA under the evaluated settings.

In conclusion, DAFSA not only provides a new interpretable and structurally grounded approach to MSA, but also marks the first successful mathematical formalization of

frame semantics. This contributes a generalizable frame-driven paradigm to multimodal semantic computation. However, the framework still faces two limitations: the dynamic-aware module retains lexical units based on a fixed ratio, which reduces adaptability to complex contextual variations, and the computational cost remains higher than that of lightweight alternatives due to multi-level attention operations. Future work will explore adaptive frame construction and model compression strategies to enhance efficiency and scalability, as well as integration with large pre-trained models for broader sentiment understanding and generation tasks.

REFERENCES

[1] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 303–311.

[2] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 108–132, 2020.

[3] R. Lin and H. Hu, "Multi-task momentum distillation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 549–565, 2024.

[4] 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*, vol. 16, no. 1, pp. 207–222, 2024.

[5] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[6] L. Chen, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "Coupled multimodal emotional feature analysis based on broad-deep fusion networks in human–robot interaction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 9663–9673, 2023.

[7] Y. Luo, J. Ma, and C. K. Yeo, "Bcmm: A novel post-based augmentation representation for early rumour detection on social media," *Pattern Recognition*, vol. 113, p. 107818, 2021.

[8] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1334–1350, 2021.

[9] V. Salaka, M. Warushavithana, N. De Silva, A. S. Perera, G. Ratnayaka, and T. Rupasinghe, "Fast approach to build an automatic sentiment annotator for legal domain using transfer learning," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 260–265.

[10] B. Biçer and H. Dibeklioğlu, "Automatic deceit detection through multimodal analysis of high-stake court-trials," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 342–356, 2023.

[11] Y. Gan, L. Xu, S. Song, and X. Tao, "Context transformer with multi-scale fusion for robust facial emotion recognition," *Pattern Recognition*, p. 111720, 2025.

[12] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 523–536, 2020.

[13] S. A. Abdu, A. H. Yousef, and A. Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Information Fusion*, vol. 76, pp. 204–226, 2021.

[14] Y. Wang, M. Liu, Z. Li, Y. Hu, X. Luo, and L. Nie, "Unlocking the power of multimodal learning for emotion recognition in conversation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5947–5955.

[15] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375–3385, 2022.

[16] D. Wang, C. Tian, X. Liang, L. Zhao, L. He, and Q. Wang, "Dual-perspective fusion network for aspect-based multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 26, pp. 4028–4038, 2023.

[17] D. Wang, Y. He, X. Liang, Y. Tian, S. Li, and L. Zhao, "Tmfn: A target-oriented multi-grained fusion network for end-to-end aspect-based multimodal sentiment analysis," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024, pp. 16 187–16 197. [Online]. Available: https://aclanthology.org/2024.lrec-main.1407/

[18] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[19] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[20] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[21] C. J. Fillmore, "Frame semantics and the nature of language," *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 20–32, 1976.

[22] S. Mai, Y. Zeng, A. Xiong, and H. Hu, "Injecting multimodal information into pre-trained language model for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2025.

[23] Y. Luo, W. Liu, Q. Sun, S. Li, J. Li, R. Wu, and X. Tang, "Triagedmsa: Triaging sentimental disagreement in multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2025.

[24] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[25] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.

[26] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[27] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2018, pp. 2225–2235.

[28] M. Li, Z. Zhu, K. Li, and H. Pei, "Diversity and balance: multimodal sentiment analysis using multimodal-prefixed and cross-modal attention," *IEEE Transactions on Affective Computing*, 2024.

[29] G. S. Krishnan, S. Padi, C. S. Greenberg, B. Ravindran, D. Manocha, and R. D. Sriram, "Linecongraphs: Line conversation graphs for effective emotion recognition using graph neural networks," *IEEE Transactions on Affective Computing*, 2025.

[30] Y. Wang, H. Jian, J. Zhuang, H. Guo, and Y. Leng, "Sslmm: Semi-supervised learning with missing modalities for multimodal sentiment analysis," *Information Fusion*, p. 103058, 2025.

[31] 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*, 2021, pp. 1009–1021.

[32] 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 conference. Association for computational linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[33] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[34] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[35] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.

[36] 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.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the*

association for computational linguistics: human language technologies, 2019, pp. 4171–4186.
[38] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 960–964.
[39] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." *SciPy*, vol. 2015, pp. 18–24, 2015.
[40] 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 (FG 2018)*, 2018, pp. 59–66.
[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[42] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 4909–4921, 2022.
[43] Z. Wu, Z. Gong, J. Koo, and J. Hirschberg, "Multimodal multi-loss fusion network for sentiment analysis," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, pp. 3588–3602.
[44] Y. Yang, X. Dong, and Y. Qiang, "Clgsi: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2099–2110.
[45] J. Huang, Y. Pu, D. Zhou, J. Cao, J. Gu, Z. Zhao, and D. Xu, "Dynamic hypergraph convolutional network for multimodal sentiment analysis," *Neurocomputing*, vol. 565, p. 126992, 2024.
[46] J. Huang, K. Jiang, Y. Pu, Z. Zhao, Q. Yang, J. Gu, and D. Xu, "Multimodal hypergraph network with contrastive learning for sentiment analysis," *Neurocomputing*, p. 129566, 2025.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[48] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115
[49] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023.
[50] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.
[51] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1424–1437, 2021.
[52] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Jian Zhuang received the B.Eng degree in computer science and technology from Qilu University of Technology, and the M.Eng degree in computer science and technology from Shandong Normal University. He is currently pursuing a Ph.D. degree in computer science and technology at Dalian University of Technology, Dalian, China. He is also a member of the Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education. His research interests include multimodal sentiment analysis and large language model.

Tiantian He (Member, IEEE) received the Ph.D. degree in computer science from The Hong Kong Polytechnic University, Hong Kong, in 2017. He is currently a Senior Scientist and Investigator with the Center for Frontier AI Research (CFAR), Institute of High Performance Computing (IHPC), Singapore Institute of Manufacturing Technology (SIMTech), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include artificial intelligence, machine learning, data-centric transfer optimization, and data mining.

Yaqing Hou received the Ph.D. degree in artificial intelligence from the Interdisciplinary Graduate School, Nanyang Technological University, Singapore, in 2017. He was a Postdoctoral Research Fellow with the Data Science and Artificial Intelligence Research Centre, Nanyang Technological University. He is currently an Associate Professor with the School of Computer Science and Technology, Dalian University of Technology, Dalian, China, and a researcher with the Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education. His research interests include computational and artificial intelligence, multi-agent learning systems, evolutionary multi-tasking, and transfer optimization. He is now an Associate Editor of the IEEE Transactions on Emerging Topics in Computational Intelligence, IEEE Transactions on Cognitive and Developmental Systems, Memetic Computing, etc. He had organized multiple Special Issues on IEEE Computational Intelligence Magazine, Memetic Computing Journal, and Applied Soft Computing.

Yew-Soon Ong (Fellow, IEEE) received the Ph.D. degree in artificial intelligence in complex design from the University of Southampton, Southampton, U.K., in 2003. He is a President's Chair Professor of Computer Science with Nanyang Technological University (NTU), Singapore, and holds the position of a Chief Artificial Intelligence Scientist of the Agency for Science, Technology and Research, Singapore. He serves as the Director of the Data Science and Artificial Intelligence Research, NTU, where he is a Co-Director of the Singtel–NTU Cognitive and Artificial Intelligence Joint Lab. His research interests are in artificial and computational intelligence. He has received several IEEE outstanding paper awards and was listed as a Thomson Reuters Highly Cited Researcher and among the World's Most Influential Scientific Minds. He is the Founding editor-in-chief of the IEEE Transactions on Emerging Topics in Computational Intelligence and an associate editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, and IEEE Transactions on Artificial Intelligence.

Qiang Zhang (Senior Member, IEEE) was born in Xi'an, China, in 1971. He received the M.Eng. degree in economic engineering and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 1999 and 2002, respectively. He was a Lecturer with the Center of Advanced Design Technology, Dalian University, Dalian, China, in 2003 and was a Professor in 2005. He is currently a Professor with the College of Computer Science and Technology, Dalian University of Technology, Dalian, China. He is the director with the Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, China. His research interests are bio-inspired computing and its applications. He has authored more than 200 papers in the above fields. Dr. Zhang has served on the editorial board of several international journals and has edited special issues in journals, such as Neurocomputing and International Journal of Computer Applications in Technology