APPENDIX

A. Details of Baseline Methods

We select a series of representative baseline methods for comparative experiments, as detailed in the following.

TFN: Tensor Fusion Network (TFN) represents the feature representations of each modality as a tensor and explicitly learns the inter-modality relationships and interactions via a triple cartesian product.

LMF: Low-rank Multimodal Fusion (LMF) enhances TFN by introducing modality-specific low-rank factors. This method reduces computational complexity while retaining essential cross-modal interaction information during the fusion process.

MulT: MulT adopts an end-to-end approach that captures the interactions between multimodal sequences through a directed pairwise cross-modal attention mechanism, without requiring explicit alignment of the data.

MISA: Modality-Invariant and -Specific Representations (MISA) effectively captures both shared and modality-specific features by learning modality-invariant and modality-specific representations within multimodal data.

Self-MM: Self-MM leverages a self-supervised label generation module, enhancing multimodal learning by strengthening the multimodal task via unimodal tasks.

CENet: CENet enhances text representations within a transformer-based pre-trained language model by integrating long-range visual and acoustic emotional cues.

MMML: MMML adopts a multi-loss training framework with cross-modal and self-attention fusion to improve multimodal sentiment representation.

CLGSI: CLGSI designs sentiment intensity-guided contrastive learning to select and weight positive and negative sample pairs, while employing a Global-Local-Fine-Knowledge fusion mechanism to extract and combine common and specific features from different modalities for sentiment intensity prediction.

Graph-MFN: Graph-MFN extends the Memory Fusion Network (MFN) by incorporating dynamic graph fusion, modeling complex inter-modality relationships using graph structures.

MTAG: Modal-Temporal Attention Graph (MTAG) introduces a novel graph construction approach and utilizes attention mechanisms to dynamically select and fuse relevant information, addressing challenges in sequence modeling.

SSLMM-CM: SSLMM-CM employs self-supervised learning to capture the structural information of the data, integrating intra- and inter-modality interactions through a graph-based alternating interaction module.

MDH: Multimodal Dynamic Hypergraph (MDH) introduces hypergraph convolutional networks into MSA for the first time, proposing a multimodal dynamic hypergraph network to learn higher-order dynamics both within and between modalities.

MHN-CL: Multimodal Hypergraph Network model based on Contrastive Learning (MHN-CL) designs supervised and unsupervised contrastive learning methods to optimize feature learning and the structure of multimodal hypergraphs based on a multimodal hypergraph neural network.

B. Implementation Details

The baseline results used in this study are derived from three implementation sources. First, the results of TFN, LMF, MulT, MISA, Self-MM, Graph-MFN, and CENet are obtained from the implementations reproduced in the open-source MMSA project on GitHub[1], using the default parameter settings for training and evaluation. Second, the results of CLGSI, MTAG, and MMML are reproduced based on the official source code released by the authors, strictly following their original training procedures and hyperparameter configurations without additional modifications. Third, for SSLMM-CM, MDH, and MHN-CL, whose source code is not publicly available, we implement the models according to the architectural design and parameter details described in the original papers, and conduct training and evaluation under a unified setting to ensure fair comparison.

In our implementation, we use the Adam optimizer with a learning-rate decay strategy, where the learning rate is reduced by a factor of 0.95 every ten iterations to accelerate convergence and improve model stability.

All training and evaluation are conducted on a workstation equipped with an AMD Ryzen Threadripper PRO 3955WX processor and an NVIDIA RTX A6000 48GB GPU. The software environment consists of Ubuntu 24.04, CUDA 12.4, Python 3.12, and PyTorch 2.5.1.

C. Detailed Comparative Results

Following the random seed setting used in the MMSA framework, we conduct the main experiments under five random seeds, i.e., [1111, 1112, 1113, 1114, 1115]. The reported results are presented as mean $\pm$ standard deviation over the five independent runs. For each evaluation metric, the strongest competing baseline is selected as the best-performing non-DAFSA method under the same experimental setting. To further examine whether the observed improvements are stable rather than caused by random initialization, we conduct statistical significance testing between DAFSA and the strongest competing baselines, with the significance level set to $p < 0.05$.

The detailed results are reported in Table I, Table II, and Table III. Overall, DAFSA maintains competitive or superior performance across the three datasets while showing relatively small standard deviations on most metrics. This indicates that the proposed frame-semantic modeling strategy not only improves the average performance, but also yields stable results across different random initializations. In particular, the consistent gains over strong graph-based and hypergraph-based baselines suggest that explicitly modeling lexical-unit evocation, semantic frame learning, and frame-level integration provides a more reliable mechanism for capturing fine-grained sentiment semantics in multimodal data.

D. Computational Cost Analysis

To evaluate whether the proposed frame-semantic modeling strategy achieves a reasonable balance between performance

[1]https://github.com/thuiar/MMSA

TABLE I

PERFORMANCE COMPARISON OF DAFSA AND BASELINE METHODS ON CMU-MOSI, REPORTED AS MEAN ± STANDARD DEVIATION. ALL COMPARATIVE RESULTS ARE SUBJECTED TO STATISTICAL SIGNIFICANCE TESTING AT $p < 0.05$. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINE INDICATES THE SECOND-BEST RESULT.

Models	Acc-2	F1-score	Acc-7	MAE	Corr
TFN	76.88/77.93 ± 1.15/1.17	76.78/77.90 ± 1.13/1.15	34.49 ± 2.78	0.968 ± 0.038	0.657 ± 0.014
LMF	78.22/79.48 ± 0.78/0.99	78.11/79.43 ± 0.68/0.87	35.60 ± 1.43	0.922 ± 0.035	0.677 ± 0.014
MulT	79.10/80.46 ± 0.86/1.05	79.08/80.50 ± 0.84/1.00	33.67 ± 0.81	0.934 ± 0.032	0.688 ± 0.008
MISA	81.98/83.72 ± 0.46/0.80	81.93/83.73 ± 0.42/0.75	42.30 ± 1.60	0.770 ± 0.020	0.776 ± 0.007
Self-MM	82.10/83.69 ± 0.60/0.59	82.04/83.68 ± 0.58/0.57	46.15 ± 1.09	**0.732** ± 0.009	0.786 ± 0.006
MMML	82.80/84.91 ± 0.87/1.03	82.71/84.89 ± 0.83/0.98	43.73 ± 2.53	0.735 ± 0.033	0.785 ± 0.010
CENet	82.13/83.84 ± 0.57/0.35	82.03/83.80 ± 0.59/0.35	43.99 ± 1.81	0.743 ± 0.017	0.783 ± 0.004
CLGSI	82.22/84.60 ± 0.93/0.97	82.05/84.53 ± 0.93/0.92	43.44 ± 0.76	0.755 ± 0.020	0.779 ± 0.021
Graph-MFN	76.79/78.02 ± 1.02/1.15	76.71/78.01 ± 1.00/1.12	34.40 ± 2.45	0.966 ± 0.051	0.653 ± 0.011
MTAG	80.76/83.08 ± 1.07/1.17	80.39/82.82 ± 1.02/1.09	25.86 ± 0.65	0.879 ± 0.036	0.706 ± 0.021
SSLMM-CM	83.38/85.09 ± 0.84/0.85	83.45/85.37 ± 0.58/0.55	46.51 ± 0.49	0.733 ± 0.007	**0.798** ± 0.005
MDH	81.02/82.47 ± 1.82/1.28	81.10/82.52 ± 1.74/1.19	38.39 ± 0.51	0.893 ± 0.031	0.694 ± 0.013
MHN-CL	82.21/84.67 ± 1.65/1.13	82.32/83.76 ± 1.64/1.10	39.45 ± 0.48	0.801 ± 0.029	0.726 ± 0.010
DAFSA	**83.94/86.28** ± 0.50/0.46	**83.71/86.14** ± 0.58/0.52	**46.75** ± 0.58	**0.732** ± 0.015	0.782 ± 0.011

TABLE II

PERFORMANCE COMPARISON OF DAFSA AND BASELINE METHODS ON CMU-MOSEI, REPORTED AS MEAN ± STANDARD DEVIATION. ALL COMPARATIVE RESULTS ARE SUBJECTED TO STATISTICAL SIGNIFICANCE TESTING AT $p < 0.05$. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINE INDICATES THE SECOND-BEST RESULT.

Models	Acc-2	F1-score	Acc-7	MAE	Corr
TFN	78.71/81.71 ± 1.48/1.33	79.12/81.54 ± 1.18/1.11	51.67 ± 0.54	0.573 ± 0.003	0.719 ± 0.003
LMF	79.14/83.47 ± 0.79/0.39	79.80/83.49 ± 0.72/0.38	52.16 ± 0.57	0.567 ± 0.006	0.734 ± 0.004
MulT	81.05/84.24 ± 1.25/0.56	81.47/84.15 ± 1.10/0.51	52.73 ± 0.65	0.560 ± 0.007	0.733 ± 0.009
MISA	80.90/84.66 ± 1.34/0.51	81.42/84.64 ± 1.10/0.49	52.72 ± 0.37	0.550 ± 0.003	0.758 ± 0.003
Self-MM	82.22/84.79 ± 1.13/0.54	82.54/84.68 ± 1.02/1.41	53.19 ± 0.53	**0.535** ± 0.003	0.766 ± 0.002
MMML	82.70/84.78 ± 1.75/1.2	82.90/84.60 ± 1.44/1.09	51.84 ± 0.43	0.553 ± 0.005	**0.768** ± 0.002
CENet	81.74/85.65 ± 0.97/0.36	82.21/85.60 ± 0.83/0.32	53.73 ± 0.50	0.539 ± 0.003	0.733 ± 0.005
CLGSI	83.24/85.11 ± 1.22/0.44	83.39/84.91 ± 1.03/0.40	53.66 ± 0.52	0.537 ± 0.006	0.763 ± 0.007
Graph-MFN	80.27/83.76 ± 1.69/0.58	80.77/83.70 ± 1.51/0.57	52.00 ± 0.20	0.564 ± 0.004	0.727 ± 0.005
MTAG	81.52/84.64 ± 1.39/0.60	81.70/84.53 ± 1.20/0.57	52.59 ± 0.56	0.559 ± 0.006	0.735 ± 0.004
SSLMM-CM	83.64/85.93 ± 0.84/0.25	83.70/86.05 ± 0.68/0.35	53.20 ± 0.40	0.551 ± 0.004	0.759 ± 0.005
MDH	82.52/84.49 ± 1.22/0.55	82.65/84.25 ± 1.06/0.49	50.41 ± 0.18	0.556 ± 0.004	0.733 ± 0.002
MHN-CL	83.31/85.95 ± 1.10/0.50	83.40/86.02 ± 1.08/0.48	51.42 ± 0.32	0.549 ± 0.004	0.745 ± 0.003
DAFSA	**84.05/86.36** ± 0.88/0.42	**84.09/86.38** ± 0.77/0.44	**53.85** ± 0.028	0.541 ± 0.003	**0.768** ± 0.002

TABLE III

PERFORMANCE COMPARISON OF DAFSA AND BASELINE METHODS ON CH-SIMS, REPORTED AS MEAN ± STANDARD DEVIATION. ALL COMPARATIVE RESULTS ARE SUBJECTED TO STATISTICAL SIGNIFICANCE TESTING AT $p < 0.05$. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINE INDICATES THE SECOND-BEST RESULT.

Model	Acc-2	Acc-3	Acc-5	F1-score	MAE	Corr
TFN	76.89 ± 0.94	64.99 ± 0.46	39.56 ± 2.67	77.22 ± 1.05	0.437 ± 0.011	0.577 ± 0.015
LMF	77.38 ± 2.16	65.38 ± 1.39	39.91 ± 2.38	77.25 ± 1.83	0.443 ± 0.012	0.567 ± 0.021
MulT	78.16 ± 1.72	65.78 ± 1.76	39.74 ± 2.03	78.10 ± 1.70	0.438 ± 0.006	0.585 ± 0.021
MISA	76.67 ± 0.83	64.07 ± 1.16	38.38 ± 2.95	76.81 ± 0.59	0.445 ± 0.006	0.566 ± 0.012
Self-MM	78.34 ± 0.74	64.82 ± 1.19	40.48 ± 1.79	78.26 ± 0.60	0.426 ± 0.012	0.584 ± 0.010
MMML	77.46 ± 0.74	65.21 ± 1.15	**44.86** ± 0.36	77.76 ± 0.67	0.431 ± 0.019	0.565 ± 0.006
CENet	77.52 ± 1.78	61.58 ± 2.65	33.25 ± 2.81	77.14 ± 1.83	0.469 ± 0.016	0.540 ± 0.039
CLGSI	78.77 ± 0.58	66.40 ± 1.05	44.42 ± 1.00	78.62 ± 0.70	0.427 ± 0.015	0.604 ± 0.010
Graph-MFN	76.93 ± 0.79	64.81 ± 0.29	40.70 ± 1.37	76.92 ± 0.60	0.438 ± 0.006	0.570 ± 0.018
MTAG	77.02 ± 0.80	54.27 ± 1.26	34.14 ± 1.29	75.59 ± 0.86	0.478 ± 0.016	0.517 ± 0.004
SSLMM-CM	80.06 ± 1.25	66.08 ± 1.69	40.87 ± 1.19	80.16 ± 0.95	0.429 ± 0.011	0.596 ± 0.009
MDH	79.21 ± 1.30	65.43 ± 1.12	41.04 ± 1.98	79.27 ± 1.03	0.422 ± 0.009	0.608 ± 0.006
MHN-CL	81.05 ± 1.23	65.68 ± 1.13	41.21 ± 1.87	**81.02** ± 1.00	0.419 ± 0.008	0.610 ± 0.007
DAFSA	**81.40** ± 0.60	**66.96** ± 0.75	**44.86** ± 1.25	80.99 ± 0.58	**0.417** ± 0.008	**0.619** ± 0.010

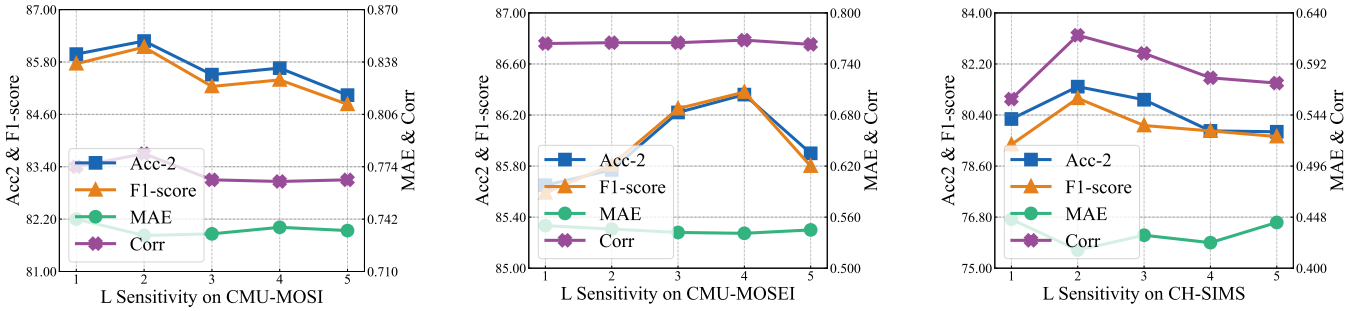

Fig. 1. Sensitivity analysis of L on different datasets.

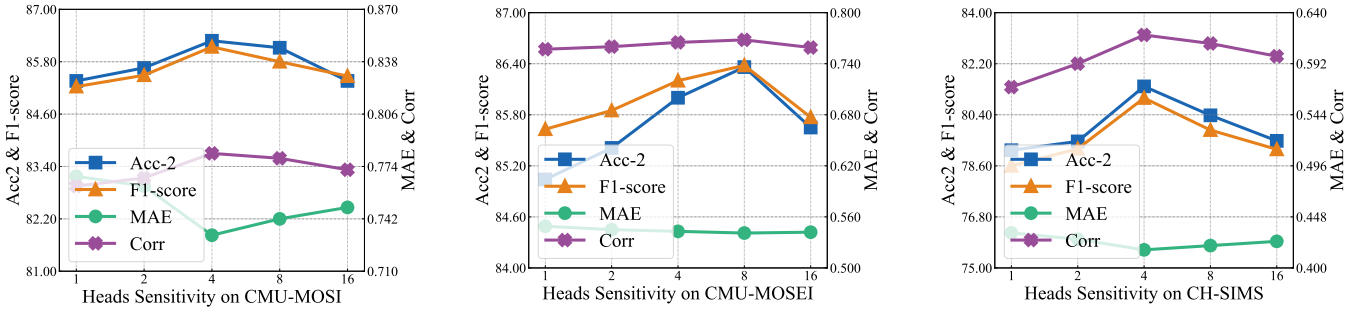

Fig. 2. Sensitivity analysis of head on different datasets

TABLE IV
COMPUTATIONAL COST (IN GFLOPS) OF DAFSA AND BASELINE MODELS ACROSS CMU-MOSI, CMU-MOSEI, AND CH-SIMS.

Model	CMU-MOSI	CMU-MOSEI	CH-SIMS
TFN	0.0321	0.0276	0.0531
LMF	0.0231	0.0231	0.0084
MulT	0.0853	0.0345	0.1809
MISA	4.2557	4.2596	3.5626
Self-MM	4.2515	4.2520	3.3221
MMML	41.3260	41.3300	27.4631
CENET	4.5476	4.5509	4.2538
CLGSI	4.2620	4.2970	3.7457
Graph-MFN	0.0950	0.0310	0.1004
MTAG	0.0032	0.0034	0.0073
SSLMM-CM	1.4768	2.2564	1.7901
MDH	0.0669	0.0720	0.1927
MHN-CL	0.1491	0.1513	0.4105
DAFSA	1.5937	2.5284	2.3133

improvement and computational cost, we compare DAFSA with several representative multimodal sentiment analysis models on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, as shown in Table IV.

As the results indicate, DAFSA requires 1.5937, 2.5284, and 2.3133 GFLOPs on the three datasets, respectively—slightly higher than lightweight models (e.g., SSLMM-CM) but significantly lower than multi-layer cross-modal attention architectures (e.g., Self-MM, CLGSI, and CENET). This suggests that DAFSA achieves a reasonable balance between performance and computational efficiency under the evaluated settings.

The main source of computational cost comes from the dynamic-aware module, where semantic frames are adaptively updated through multi-level attention and message propagation. While this process introduces additional computation, it effectively enhances fine-grained semantic representation and model robustness. Future work will explore lightweight strategies such as pruning and knowledge distillation to further reduce computational overhead without sacrificing performance.

E. Parameter Sensitivity

In this section, we conduct parameter sensitivity experiments to investigate the robustness of DAFSA with respect to key hyperparameters. Specifically, we examine two hyperparameters closely related to frame-semantic modeling: the number of layers L in the SFPU, which affects the depth of frame learning and integration; and the number of attention heads, which influences multi-subspace frame interaction.

1) Sensitivity Analysis of L: To analyze the impact of the number of SFPU layers L on model performance, we conducted experiments with $L = \{1, 2, 3, 4, 5\}$ on different datasets. The experimental results, as shown in Fig. 1, reveal that optimal performance is achieved with L = 2, 4, and 2 on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, respectively. Shallower networks exhibit the best performance on smaller datasets such as CMU-MOSI and CH-SIMS. This phenomenon suggests that, in cases of smaller or less complex datasets, shallow networks can effectively avoid overfitting and rapidly capture the core sentiment features within the sentiment information. However, on larger datasets like CMU-MOSEI, deeper networks can fully utilize their larger model capacity to learn complex sentiment patterns in multimodal information, thereby achieving optimal performance. This further validates the adaptability between network depth and dataset size: on larger and more complex datasets, deeper

networks can effectively process and integrate multi-level sentiment features, while on smaller datasets, shallower networks better avoid overfitting and improve the model's generalization ability.

2) Sensitivity Analysis of Heads: To further investigate the impact of the number of attention heads on model performance, we conducted systematic experiments on datasets of different scales. The results, as shown in Fig. 2, indicate that when the number of attention heads is set to 4, 8, and 4, the model achieves optimal performance on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, respectively. This trend reveals that, in small-scale datasets (such as CMU-MOSI and CH-SIMS), a moderate number of attention heads effectively capture the key attention relationships between tokens and the semantic frame, avoiding information dilution caused by excessive attention dispersion. On the larger and more complex CMU-MOSEI dataset, more attention heads help in parallel modeling of fine-grained sentiment features across different semantic dimensions, enhancing the model's ability to perceive diverse sentiment expressions. However, as the number of attention heads increases further, the model performance generally decreases, indicating that the number of attention heads should be aligned with the scale and complexity of the dataset to avoid information redundancy or overfitting.

F. Semantic Frame Visualization for CH-SIMS

In addition, we further select a Chinese sample from the CH-SIMS test set for visualization analysis. As shown in Fig. 3, for the utterance "现在好了，汽水不冰了，可我的心却是冰冰的", DAFSA organizes the textual lexical units into three main semantic frames. Specifically, "汽/ 水/ 不/ 冰/ 了" forms a state-change frame, describing the external state that the soda is no longer cold; "好/ 了/ 可/ 却/ 心" forms a contrast frame, indicating the semantic contrast between the preceding and following clauses; and "心/ 却/ 是/ 冰/ 冰" forms a negative-emotion frame, expressing the speaker's cold or disappointed inner state.

Overall, the visualization results provide qualitative evidence that DAFSA can be applied to both Chinese and English datasets, showing the cross-language applicability of the proposed frame-semantic modeling strategy.

G. Case Study and Error Analysis

To further clarify the applicability boundaries of DAFSA, we conduct a qualitative case study and error analysis on the CH-SIMS test set. CH-SIMS is selected because it provides both multimodal sentiment labels and modality-specific sentiment labels, which allows us to examine whether model predictions are affected by modality consistency or modality conflict. It should be noted that the modality-specific labels are used only for qualitative analysis and are not used for training DAFSA.

Table V presents two representative correct predictions and two incorrect predictions from the CH-SIMS test set. The correct cases show that DAFSA tends to produce reliable predictions when textual, acoustic, and visual cues provide consistent or complementary sentiment evidence. For example,

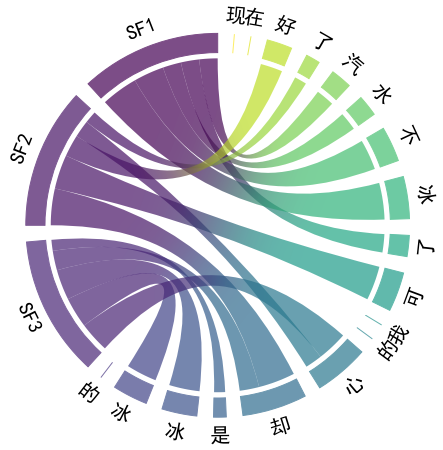

Fig. 3. Semantic frame visualization of a Chinese sample from CH-SIMS.

TABLE V
QUALITATIVE ANALYSIS OF REPRESENTATIVE CORRECT AND INCORRECT PREDICTIONS ON CH-SIMS.

Utterance	Modality Labels	GT	Pred.
现在好了，汽水不冰了，可我的心却是冰冰的	T: -1.0, A: -0.6, V: -1.0	-1.0	-0.86
但是我还是来了，我觉得我站在这就是赢了，对我自己来讲	T: 1.0, A: 1.0, V: 0.8	0.8	0.86
张扬，谢谢你	T: 1.0, A: 1.0, V: 0.8	-0.6	0.35
我整个人裂开了	T: 1.0, A: 1.0, V: 0.8	-0.2	0.69

in the utterance one, all modality-specific labels indicate negative sentiment. Although the first part of the utterance describes an apparent improvement in the external situation, the contrastive expressions shift the sentiment focus to the speaker's inner emotional state, leading to a negative sentiment prediction. In contrast, the incorrect cases reveal the applicability boundaries of DAFSA. When the modality-specific labels are inconsistent with the final multimodal label, or when the utterance relies heavily on external discourse context, sarcasm, or scene-level background, the model may have difficulty determining the dominant sentiment evidence. In such cases, lexical-frame activation based on the current utterance may be insufficient to recover the intended sentiment, leading to biased predictions. Overall, DAFSA is more suitable for samples where multimodal cues provide coherent or complementary semantic evidence and where sentiment semantics can be organized through lexical units and their evoked frames. These observations further clarify the applicability boundaries of the proposed method.