Hierarchical Global-Local Interaction and Refinement for Multimodal Sentiment Analysis

Yuanyuan Zhou , Yanzhang He, Anni Wang, and Yu Lu(✉)

Sichuan Normal University, Chengdu, China
20251393038@stu.sicnu.edu.cn

Abstract. Multimodal Sentiment Analysis (MSA) aims to integrate text, audio, and visual modalities to achieve accurate sentiment modeling. Existing methods often rely on shallow interaction structures, making it difficult to jointly capture fine-grained local dynamics and high-level semantic dependencies. In addition, dominant modalities may suppress the learning of weaker modalities, leading to insufficient cross-modal semantic alignment. To address these issues, we propose a Hierarchical Global-Local Interaction and Refinement framework for Multimodal Sentiment Analysis (HGLIR). Specifically, a Modality Dropout strategy is first introduced at the input stage to alleviate over-reliance on a single modality and improve robustness. Based on this, a Hierarchical Local Interaction (HLI) module models multimodal sequences through a multi-layer progressive structure to capture local dynamic features at different semantic levels. Within the HLI module, a Cross-Modal Synergistic Learning (CMSL) mechanism explicitly models cross-modal semantic consistency and gradually aligns information during interaction. Furthermore, a Global Representation Refinement (GRR) module introduces learnable global representations and iteratively updates them in a multi-layer structure to aggregate long-range semantic dependencies and form stable high-level semantic representations. Experimental results on CMU-MOSI and CMU-MOSEI demonstrate the effectiveness of the proposed framework across multiple evaluation metrics.

Keywords: Multimodal Sentiment Analysis, Feature Fusion, Attention Mechanism, Hierarchical Interaction

1 Introduction

Multimodal Sentiment Analysis (MSA) has emerged as an important research topic in computer vision and natural language processing [1][2]. In real-world scenarios, human emotions are usually expressed through multiple signals, including text, audio, and visual information. Effectively integrating information from different modalities and exploiting their complementarity have become key challenges in MSA [3].

Recent research has increasingly emphasized the development of more effective representation learning methods for MSA. Early methods mainly relied on feature concatenation or simple fusion strategies, which were insufficient for modeling cross-modal dependencies effectively [1]. Subsequently, a series of methods based on attention

mechanisms and Transformer architectures were proposed. These methods employed cross-modal attention to achieve information interaction and significantly improved model performance. Additionally, some studies further introduced multi-layer interaction or graph-based modeling to capture more complex semantic relationships. However, most existing approaches primarily focus on modeling information at a single semantic level. They either emphasize local temporal dynamics while ignoring global semantic dependencies, or rely heavily on global representations without sufficiently modeling fine-grained interactions.

To address these issues, this paper proposes a hierarchical global-local interaction and refinement framework for multimodal sentiment analysis.

The main contributions of this paper are summarized as follows:

(1) We propose a Hierarchical Local Interaction (HLI) module and introduce a Cross-Modal Synergistic Learning (CMSL) mechanism into the multi-layer interaction process, which achieves gradual alignment and fusion of multimodal sequences, thereby enhancing local dynamic representation learning.

(2) We design a Global Representation Refinement (GRR) module. By introducing learnable global representations and combining them with cross-modal attention, the module performs iterative updates in a multi-layer structure, thereby effectively capturing long-range dependencies and high-level semantic structures.

(3) Extensive experiments and ablation analyses are performed on the CMU-MOSI and CMU-MOSEI datasets. The results indicate that HGLIR outperforms existing approaches across multiple evaluation metrics, validating the effectiveness of the proposed framework for multi-level semantic modeling and cross-modal interaction.

2 Related Work

2.1 Multimodal Sentiment Analysis

Early MSA approaches primarily concentrated on multimodal fusion strategies. TFN [4] models high-order interactions through tensor fusion, while LMF [5] reduces computational complexity via low-rank factorization. Later studies shifted toward representation learning and semantic alignment. MISA [6] disentangles shared and private representations to reduce modality discrepancy, and Self-MM [7] introduces self-supervised multi-task learning for cross-modal consistency modeling. Transformer-based methods, such as MulT [3] and MAG-BERT [8], further improve cross-modal interaction through attention mechanisms. However, most existing approaches still depend on shallow or single-level fusion architectures, making it difficult to jointly capture local dynamics and global semantic information. In addition, dominant text features often suppress non-linguistic modalities, limiting adaptive multimodal fusion and model generalization [1][2].

2.2 Cross-Modal Hierarchical Representation Learning

To better capture cross-modal semantic relationships, recent studies have gradually evolved from information-constrained fusion methods to structured hierarchical

modeling, and further explore collaborative representation learning between global and local information. Early work such as MMIM [9] introduces mutual information maximization constraints at different levels from an information-theoretic perspective, enhancing information sharing among modalities and improving the preservation of task-relevant information in fused representations.

On this basis, later studies began to introduce explicit structures for modeling multi-level semantic relationships. For example, HGCL-LG [10] constructs local and global structures and applies a hierarchical learning to model cross-modal relationships at different granularities. Furthermore, GloMo [11] separates modality representations into local and global components and integrates them through a global guidance mechanism, thereby improving fine-grained sentiment modeling capability.

Inspired by these studies, we argue that an effective multimodal sentiment analysis model should simultaneously capture local semantic details and global semantics, and enable adaptive information integration across different semantic levels. However, existing methods often fail to achieve this balance. Therefore, we propose a hierarchical global-local interaction and refinement model, which progressively models cross-modal dynamics through multi-layer local interaction and captures high-level semantic dependencies via an iterative global representation refinement mechanism. Furthermore, adaptive fusion is introduced to achieve unified modeling of multi-scale information.

3 Method

3.1 Overview

As shown in Fig.1, the proposed HGLIR framework first obtains aligned multimodal representations through modality-specific encoders and modality dropout. The representations are then input into the Hierarchical Local Interaction (HLI) module and the Global Representation Refinement (GRR) to capture local dynamics and global semantic dependencies, respectively. Finally, global and local representations are adaptively fused for sentiment prediction.

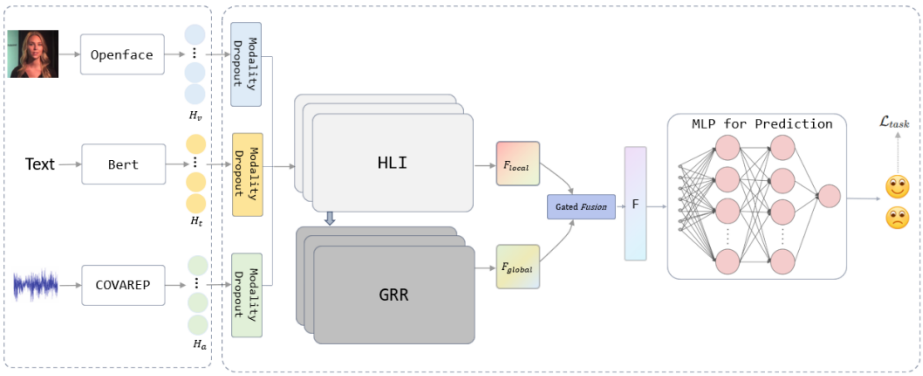


Fig. 1. HGLIR framework architecture

3.2 Unimodal Feature Representation

In multimodal sentiment analysis tasks, each video clip typically consists of signals from three different modalities: text t, audio a, and visual v. Following common practices in existing research, feature extraction is first performed on the data of each modality. Specifically, the textual modality is encoded by BERT [12] to obtain contextual semantic representations. The audio and visual modalities are processed using COVAREP [13] and OpenFace [14], respectively, to extract the corresponding low-level features.

After obtaining these features, the input sequences of the three modalities are uniformly represented as (X_t, X_a, X_v), where $X_m \in \mathbb{R}^{T_m \times d_m}$, $m \in \{t, a, v\}$. Here, T_m denotes the sequence length of each modality m, and d_m represents the corresponding feature dimension. To facilitate subsequent cross-modal interaction modeling, a modality-specific linear projection is further applied to project the features into multimodal representations of the same dimension $H_m \in \mathbb{R}^{T_m \times d}$.

3.3 Modality Dropout

The quality and availability of multimodal signals are often uneven, causing certain modalities to dominate the learning process. This imbalance may lead the model to rely excessively on a single modality, thereby weakening its generalization capability in scenarios with missing or noisy modalities [15]. Consequently, we introduce Modality Dropout, a modality-level random dropout mechanism designed to improve model robustness and generalization capability. During training, a random mask is applied to each modality feature H_m to randomly discard the entire modality for each sample, which can be formally defined as follows:

$$\widetilde{H}_m = \frac{M_m}{1 - p_m} \cdot H_m, m \in \{t, a, v\} \tag{1}$$

where $M_m \sim \text{Bernoulli}(1 - p_m)$ denotes the sample-level random mask. When $M_m = 0$, the corresponding modality is completely discarded. When $M_m = 1$, the feature is scaled to maintain expectation consistency.

Different dropout probabilities (p_t, p_a, p_v) are assigned to different modalities. To alleviate over-reliance on the text modality, a higher dropout probability is applied to text, while lower probabilities are used for audio and visual modalities.

3.4 Hierarchical Local Interaction

Effective multimodal understanding requires progressively modeling semantic interactions at different representation levels. However, existing methods usually rely on shallow interaction structures, hindering the ability to jointly model local dynamics and high-level semantic relationships, thereby limiting multi-scale semantic modeling capability [16].

To tackle this issue, we propose a Hierarchical Local Interaction (HLI) module with multi-layer cross-modal interaction. Meanwhile, a Cross-Modal Synergistic Learning

(CMSL) mechanism is introduced at each layer to progressively align cross-modal semantics. The interaction process is shown in Fig. 2.

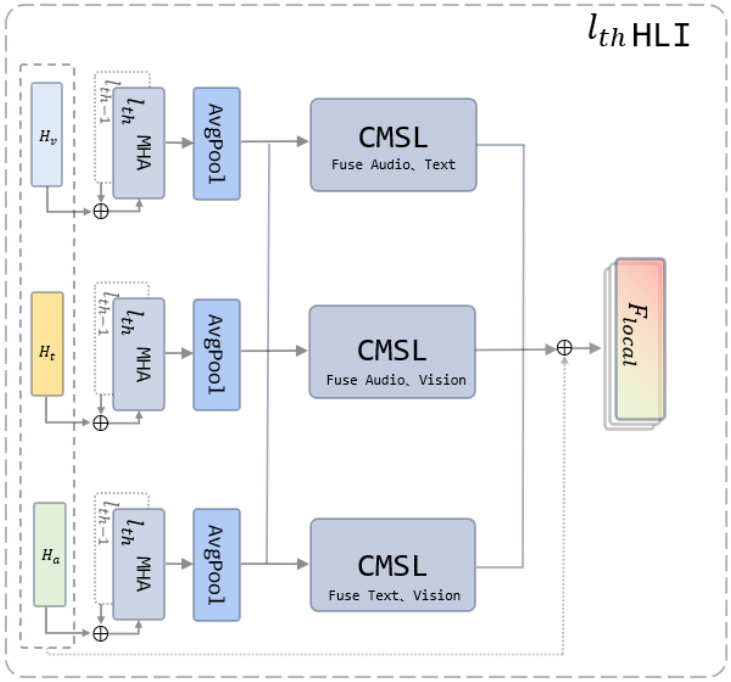


Fig. 2. Framework of the HLI module

Specifically, the multimodal representation of the l-th layer is updated layer by layer from the output of the previous layer:

$$\left(H_t^{(l)}, H_a^{(l)}, H_v^{(l)}\right) = \mathrm{HLI}\left(H_t^{(l-1)}, H_a^{(l-1)}, H_v^{(l-1)}\right), l = 1, 2, \ldots, L \tag{2}$$

where $H_m^{(l)} \in \mathbb{R}^{T_m \times d}$ denotes the sequence representation of modality m at the l-th layer. L is set to 3 by default. The initial representation is obtained from the modality dropout process, i.e., $\mathrm{H}_{\mathrm{m}}^{(0)} = \widetilde{H}_m$.

At each layer, temporal modeling is first performed independently for each modality to capture intra-modal dynamic dependencies. Specifically, we apply a Multi-Head Attention (MHA) mechanism [17], followed by a residual connection to obtain the updated representation:

$$\widehat{H}_m^{(l)} = H_m^{(l-1)} + MHA_m\left(H_m^{(l-1)}\right) \tag{3}$$

Relying solely on attention mechanisms for implicit alignment often fails to effectively constrain semantic consistency across modalities. This can easily lead to an unstable cross-modal interaction process. To address this, a Cross-Modal Synergistic Learning (CMSL) mechanism is designed to optimize the cross-modal interaction.

Specifically, an average pooling (AvgPool) is first performed on the updated sequences of each modality:

$$f_m^{(l)} = AvgPool\left(\widehat{H}_m^{(l)}\right) \tag{4}$$

Cosine similarity is adopted to measure semantic consistency across modalities:

$$S_{ij}^{(l)} = \sigma\left(\cos\left(f_i^{(l)}, f_j^{(l)}\right)\right), i \neq j \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid function.

Leveraging these similarity weights, we dynamically fuse cross-modal information and subsequently apply a residual connection to yield the final representation:

$$\widetilde{H}_m^{(l)} = \widehat{H}_m^{(l)} + \alpha \sum S_{mn}^{(l)} \cdot \widehat{H}_n^{(l)},\ n \neq m \tag{6}$$

where α is a learnable parameter, and $S_{mn}^{(l)}$ represents the semantic similarity weight between modality m and n.

After generating the final local representation for each modality through multi-layer progressive modeling, we fuse these multimodal features to extract the local semantic features:

$$F_{local} = Concat\left(\widetilde{H}_t^{(L)}, \widetilde{H}_a^{(L)}, \widetilde{H}_v^{(L)}\right) W_f \tag{7}$$

where W_f is a learnable linear projection, and L is the last layer of the hierarchical interaction (i.e., the final hierarchical representation obtained after multi-layer progressive modeling).

3.5 Global Representation Refinement

Although the HLI module can effectively capture cross-modal dynamic information, its modeling process is still limited to a local temporal range. This makes it difficult to fully characterize global semantic dependencies. To this end, we introduce a Global Representation Refinement (GRR) module. By introducing learnable global representations and continuously aggregating information from different modalities and hierarchical layers throughout the multi-layer interaction process, the model progressively captures and refines global semantics. The refinement process is shown in Fig. 3.

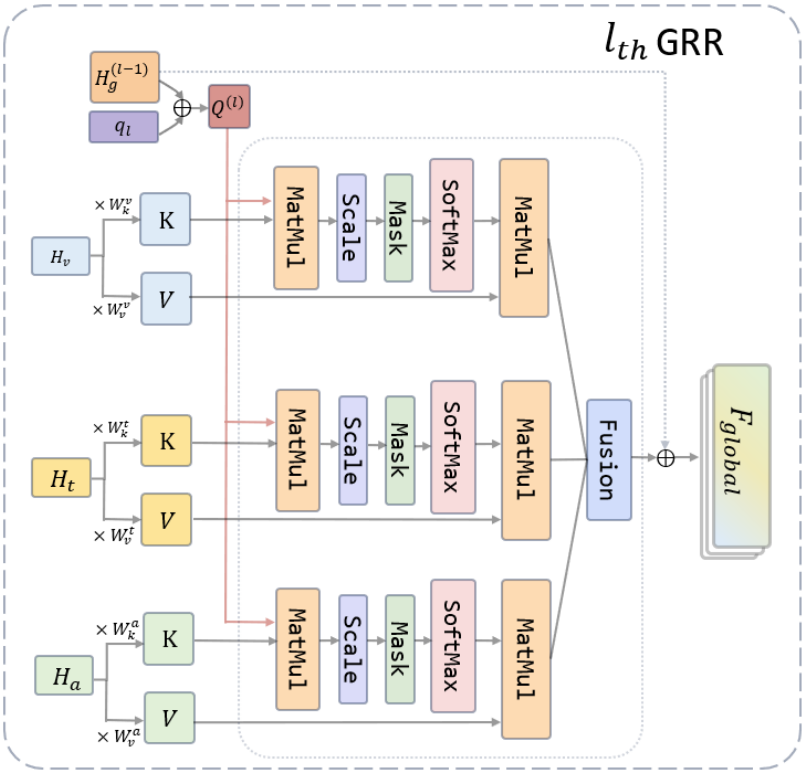


Fig. 3. Framework of the GRR module

Specifically, we first introduce a set of randomly initialized learnable global representations as the initial representation for semantic aggregation, denoted as $H_g^{(0)} \in \mathbb{R}^{T\times d}$, where T represents the global sequence length and d is the feature dimension. In the update, we construct the query vector by applying a linear projection to the global representation from the previous layer, and introduce an additional learnable query vector to enhance the representation capability:

$$Q^{(l)} = \left(H_g^{(l-1)} + q_l\right) W_q \tag{8}$$

where $W_q \in \mathbb{R}^{d\times d_q}$ represents the learnable query projection matrix, and q_l is a learnable parameter used to enhance query expressiveness.

For the construction of keys and values, we utilize the representations of each modality to generate the corresponding keys and values:

$$K_m = H_m W_k^m, V_m = H_m W_v^m, m \in \{t, a, v\} \tag{9}$$

where W_k^m and W_v^m are learnable linear transformation parameters. H_m denotes the final modality representation obtained after the multi-layer modeling of the Hierarchical Local Interaction (HLI) module, i.e., $H_m = \widetilde{H}_m^{(L)}$.

Subsequently, the attention mechanism is employed to extract globally relevant information from each modality.

$$Z_m^{(l)} = \text{Softmax}\left(\frac{Q^{(l)}K_m^T}{\sqrt{d}}\right)V_m, m \in \{t, a, v\} \tag{10}$$

The attention outputs from different modalities are then fused to obtain the updated global representation through a linear projection:

$$\widehat{H}_g^{(l)} = \left(Z_t^{(l)} + Z_v^{(l)} + Z_a^{(l)}\right) W_o \tag{11}$$

where W_o represents the learnable linear projection parameter.

Finally, the update is completed via a residual connection:

$$H_g^{(l)} = H_g^{(l-1)} + \widehat{H}_g^{(l)} \tag{12}$$

This process executes iteratively across multiple levels. Through layer-by-layer updates, the global representation can progressively integrate multimodal information to form a stable and consistent high-level semantic representation, ultimately obtaining the final global representation $F_{\text{global}} = H_g^{(L)}$.

3.6 Global-Local Fusion Module

Local representations focus on depicting fine-grained semantic information formed during the multi-layer interaction process, whereas global representations aggregate multimodal information at a macro level. The two are complementary in semantic granularity. Since the relative importance of local and global information varies across different samples, introducing a gating mechanism enables adaptive information fusion.

Specifically, a gating mechanism is applied to the local and global representations to adaptively regulate the contribution of information from different sources based on the input content. The process is defined as:

$$g_l = \sigma(F_{\text{local}} W_l), g_g = \sigma\left(F_{\text{global}} W_g\right) \tag{13}$$

$$F = g_l \odot \mathrm{F}_{\text{local}} + g_g \odot \mathrm{F}_{\text{global}} \tag{14}$$

where W_l and W_g are learnable linear transformation parameters. g_l and g_g represent the gating weights for the local and global branches, respectively.

Through the above process, the model can simultaneously integrate multi-layer local semantics, global semantics, and modality-level importance information within a unified representation space. This results in a more comprehensive and discriminative multimodal representation.

3.7 Prediction and Training Objective

Prediction. After obtaining the global-local fused representation F, it is input into the prediction layer to complete the final prediction task. Specifically, we use a Multi-Layer Perceptron (MLP) as the prediction function to map the fused representation:

$$\hat{y} = \mathrm{MLP}(F; \theta^{\mathrm{MLP}}) \tag{15}$$

where $\hat{y}$ denotes the sentiment result predicted by the model. $\mathrm{MLP}(\cdot)$ denotes a multi-layer prediction network composed of two linear transformation layers with a ReLU activation function, and θ^{MLP} represents its set of learnable parameters.

Training Objective. For the MSA task, the Mean Squared Error (MSE) loss is adopted as the optimization objective:

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{16}$$

4 Experiments

4.1 Datasets

The proposed method is evaluated on two benchmark datasets, CMU-MOSI [18] and CMU-MOSEI [19]. CMU-MOSI contains 2,199 video clips from 93 speakers, while CMU-MOSEI is a larger-scale dataset with over 23,000 sentence-level clips collected from more than 1,000 speakers across 250 topics. Both datasets provide aligned text, audio, and visual modalities.

4.2 Evaluation Metrics

We refer to existing studies and adopt five general evaluation metrics to comprehensively evaluate the overall performance of the model. These metrics include Mean Absolute Error (MAE), Pearson Correlation Coefficient (Corr), binary classification accuracy (Acc2), 7-class Accuracy (Acc7), and F1-score (F1). Higher values indicate better performance for all metrics except MAE.

4.3 Experimental Settings

The experiments are conducted on the Ubuntu 18.04 operating system environment. We implement and train the model using PyTorch 2.6 and Python 3.11. The hardware consists of an NVIDIA GeForce RTX 4090 GPU. The hyperparameters used in the experiments are shown in Table 1.

Table 1. Experimental Settings

Parameter	CMU-MOSI	CMU-MOSEI
learning rate	1e-4	1e-4
batch size	64	64
Optimizer	Adam	Adam
Epochs	100	100
Multi-head attention heads	8	8
p_t	0.15	0.15
p_a	0.10	0.10
p_v	0.10	0.10
Local layers	3	3
Global layers	3	3

4.4 Baseline Models

To comprehensively evaluate the proposed model, we adopt the following classic and cutting-edge models as baselines for MSA.

Early methods mainly centered on feature fusion and cross-modal interaction modeling. Representative approaches include TFN [4] and LMF [5], which model multimodal feature interactions through tensor fusion and low-rank decomposition, respectively. Attention-based and pre-trained architecture methods, such as MulT [3] and MAG-BERT [8], further improved cross-modal alignment and interaction capabilities by introducing directional cross-modal attention and multimodal adaptation mechanisms. To enhance representation quality, methods including MISA [6] and Self-MM [7] adopted disentanglement learning and self-supervised strategies to learn robust multimodal representations. More recent studies, such as MMIM [9], HGCL-LG [10], GLoMo [11], TMBL [20], DLF [21], and EMOE [22], further explored hierarchical semantic modeling, modality balancing, language-focused fusion, and dynamic expert mechanisms, leading to stronger multimodal representation and sentiment prediction performance.

4.5 Experimental Results

On two public benchmark datasets, we systematically compare the proposed model with various multimodal sentiment analysis baselines to evaluate the effectiveness of the proposed model. The results are shown in Table 2 and Table 3.

Table 2. Experimental Results on CMU-MOSI

Model	Corr	MAE	Acc7	Acc2	F1
TFN[1]	0.673	0.947	34.4	77.9/79.0	77.9/79.1
LMF[1]	0.651	0.950	33.8	77.9/79.1	77.8/79.1
MulT[1]	0.725	0.846	40.4	79.71/83.4	79.6/83.5
MISA[1]	0.778	0.776	41.3	81.8/83.5	81.8/83.5
MAG-BERT[3]	0.789	0.731	-	82.5/84.3	82.6/84.3
Self-MM[1]	0.796	0.708	46.6	83.4/85.4	83.3/85.4
MMIM[2]	0.800	0.700	46.6	84.1/86.0	84.0/85.9
HGCL-LG[2]	0.788	0.725	41.7	-/84.0	-/83.9
GLoMo[2]	0.782	0.718	**48.3**	84.1/**86.7**	83.9/**86.6**
TMBL[2]	0.762	0.867	36.3	81.7/83.8	82.4/84.2
DLF[2]	0.781	0.731	47.0	–/85.0	–/85.0
EMOE[2]	-	0.710	47.8	-/85.4	-/85.3
OURS	**0.802**	**0.687**	48.2	**84.4**/**86.7**	**84.2**/86.5

Note. The best results are indicated in **bold**. [1] indicates that the data is obtained from THUIAR's GitHub page, [2] indicates that the data is from the original paper, and [3] indicates that the data is from [9]. The dash ("-") represents missing results for the corresponding metric in the original paper. For the F1 and Acc2 metrics, the left side of "/" indicates negative/positive results, while the right side indicates negative/non-negative results.

Table 3. Experimental Results on CMU-MOSEI

Model	Corr	MAE	Acc7	Acc2	F1
TFN[1]	0.714	0.572	51.6	78.5/81.8	78.9/81.7
LMF[1]	0.716	0.575	51.5	80.5/83.4	80.9/83.3
MulT[1]	0.733	0.559	52.8	81.1/84.6	81.5/83.6
MISA[1]	0.751	0.557	52.0	80.6/84.6	81.1/84.6
MAG-BERT[3]	0.753	0.539	-	83.8/85.2	83.7/85.1
Self-MM[1]	0.764	0.530	45.8	83.7/85.1	83.8/84.9
MMIM[2]	0.772	**0.526**	54.2	82.2/85.9	82.6/85.9
HGCL-LG[2]	0.769	0.545	49.3	-/84.2	-/84.3
GLoMo[2]	0.771	0.539	55.0	83.7/**86.5**	84.0/**86.4**
TMBL[2]	0.766	0.545	52.4	**84.2**/85.8	**84.8**/85.9
DLF[2]	0.764	0.536	53.9	–/85.4	–/85.2
EMOE[2]	-	0.536	54.1	-/85.3	-/85.3
OURS	**0.775**	0.537	**55.3**	83.9/86.2	83.7/86.1

On the CMU-MOSI dataset, our method performs well in the regression task, achieving current competitive results with Corr = 0.802 and MAE = 0.687. Compared to the strong baseline MMIM (Corr = 0.800, MAE = 0.700), the correlation coefficient increases by 0.002, and the MAE decreases by 0.013. Meanwhile, compared to Self-MM (MAE = 0.708), the error is further reduced by 0.021, which demonstrates that the model possesses a stronger fitting capability in sentiment intensity modeling. In the fine-grained sentiment classification task (Acc7), our method achieves a score of

48.2%, which is on par with the current top-performing method GLoMo (48.3%) and outperforms EMOE (47.8%) by approximately 0.4%, showing strong multi-class discriminative ability. In the binary classification task, the model achieves an Acc2 of 84.4% / 86.7% and an F1-score of 84.2% / 86.5%. These results remain consistent with existing methods, reflecting good classification stability.

On the CMU-MOSEI dataset, our method achieves a score of 55.3% in the fine-grained sentiment classification task (Acc7). This result outperforms the current top-performing method GLoMo (55.0%) by 0.3% and EMOE (54.1%) by 1.2%, reaching the competitive level and further verifying the effectiveness of the model in fine-grained sentiment modeling. Regarding the regression metrics, the model achieves MAE = 0.537 and Corr = 0.775, showing comparable performance with advanced methods such as MMIM (MAE = 0.526) and Self-MM (MAE = 0.530).

Compared with hierarchical methods such as MMIM, HGCL-LG, and GLoMo, our model achieves competitive and balanced performance across multiple evaluation metrics. In particular, it shows stronger regression capability while maintaining favorable classification performance, demonstrating the effectiveness of hierarchical global-local modeling for multimodal sentiment analysis.

4.6 Ablation Study

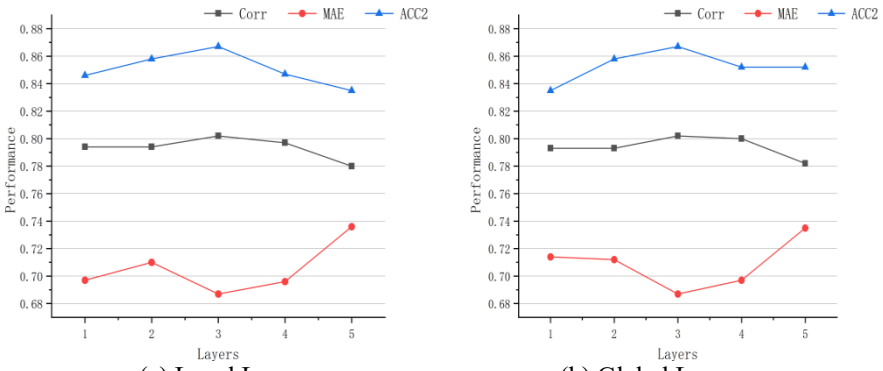
Ablation Analysis of Key Components. To comprehensively evaluate the contribution of each key module to the model's performance, we design a series of ablation experiments. Specifically, we remove core components from the model one at a time, including the Modality Dropout mechanism, the Hierarchical Local Interaction (HLI) module, and the Global Representation Refinement (GRR) module. We also consider removing both HLI and GRR simultaneously to analyze the synergistic effect between local and global modeling. When a module is removed, the model architecture is appropriately simplified to ensure computational correctness. The results are shown in Table 4.

Table 4. Ablation Results on CMU-MOSI

Model	Corr	MAE	Acc7	Acc2	F1
Base	**0.802**	**0.687**	**48.2**	**86.7**	**86.5**
w/o Modality Dropout	0.787	0.711	47.0	86.4	86.3
w/o HLI	0.772	0.736	43.8	84.4	84.2
w/o GRR	0.780	0.732	46.5	85.9	85.8
w/o HLI&GRR	0.751	0.747	43.1	83.8	83.7

The removal of any component leads to performance degradation. Removing HLI causes the largest drop, indicating the importance of hierarchical local interaction for modeling multimodal dynamics. Removing GRR also reduces performance, validating the effectiveness of global semantic refinement. The worst performance is observed when both HLI and GRR are removed, demonstrating their complementarity. In addition, Modality Dropout further improves model robustness.

Analysis of Layer Settings. To verify the rationality of the layer configurations within the hierarchical architecture, we conduct comparative experiments on the number of layers for both the local interaction module (Local Layers) and the global representation refinement (Global Layers), evaluating the performance variations on the CMU-MOSI. The results of the experiments are illustrated in Fig. 4.



(a) Local Layers. (b) Global Layers.

Fig. 4. Global and Local Layer number analysis on CMU-MOSI. For ease of presentation and comparison, only three metrics—Corr, MAE, and Acc2—are selected in the figure, and the values of Acc2 are divided by 100 for normalization.

Experimental results show that both the local interaction layers (Local Layers) and the global representation refinement layers (Global Layers) exhibit a similar trend, where the model performance first improves and then declines when the number of layers increases. Specifically, the performance consistently improves when the number of layers increases from 1 to 3, indicating that multi-layer local interactions and global updates can effectively capture richer cross-modal dynamics and aggregate long-range semantic information. However, further increasing the number of layers to 4 or 5 leads to performance fluctuations or degradation, which may be caused by redundant information, optimization difficulties, or over-smoothing effects introduced by excessively deep structures. Therefore, both the local interaction and global representation refinement are set to 3 layers.

5 Conclusion

This paper proposes HGLIR for multimodal sentiment analysis, which jointly models hierarchical local interaction and global semantic refinement to capture multi-scale semantic information. Specifically, the proposed framework progressively learns cross-modal dynamics through the Hierarchical Local Interaction (HLI) module and enhances semantic consistency via the Cross-Modal Synergistic Learning (CMSL) mechanism. Meanwhile, the Global Representation Refinement (GRR) module iteratively aggregates long-range semantic dependencies through learnable global representations.

Experimental results on the CMU-MOSI and CMU-MOSEI datasets demonstrate that the proposed method achieves competitive performance across multiple metrics, further validating the effectiveness of hierarchical global-local modeling for multimodal sentiment analysis.

References

1. Zhu L, Zhu Z, Zhang C, et al. Multimodal sentiment analysis based on fusion methods: A survey[J]. Information Fusion, 2023, 95: 306-325.
2. Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.
3. Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6558-6569
4. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor fusion network for multimodal sentiment analysis. In: EMNLP, pp. 1103–1114 (2017)
5. Liu, Z., Shen, Y.: Efficient low-rank multimodal fusion with modality-specific fac-tors. In: ACL, pp. 2247–2256 (2018)
6. Hazarika D, Zimmermann R, Poria S. Misa: Modality-invariant and-specific representations for multimodal senti-ment analysis [C]//Proceedings of the 28th ACM internat-ional conference on multimedia.2020: 1122-1131.
7. Yu W, Xu H, Yuan Z, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 10790-10797.
8. Rahman, W., et al.: Integrating multimodal information in large pretrained transformers. In: Proceedings of the Conference. Association f or Computational Lin guistics. Meeting, vol. 2020, p. 2359. NIH Public Access (2020)
9. Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[C]//Proceedings of the 2021 conference on empirical methods in natural language processing. 2021: 9180-9192.
10. Du J, Jin J, Zhuang J, et al. Hierarchical graph contrastive learning of local and global presentation for multimodal sentiment analysis[J]. Scientific Reports, 2024, 14(1): 5335.
11. Zhuang Y, Zhang Y, Hu Z, et al. Glomo: Global-local modal fusion for multimodal sentiment analysis[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 1800-1809.
12. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
13. Degottex G, Kane J, Drugman T, et al. COVAREP—A collaborative voice analysis repository for speech technologies[C]//2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2014: 960-964.
14. Baltrušaitis T, Robinson P, Morency L P. Openface: an open source facial behavior analysis toolkit[C]//2016 IEEE winter conference on applications of computer vision (WACV). IEEE, 2016: 1-10.

15. Wang W, Tran D, Feiszli M. What makes training multi-modal classification networks hard?[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12695-12705.
16. Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 5100-5111.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
18. Zadeh A, Zellers R, Pincus E, et al. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.
19. Zadeh A A B, Liang P P, Poria S, et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2236-2246.
20. Huang J, Zhou J, Tang Z, et al. TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis[J]. Knowledge-Based Systems, 2024, 285: 111346.
21. Wang P, Zhou Q, Wu Y, et al. DLF: Disentangled-language-focused multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(20): 21180-21188.
22. Fang Y, Huang W, Wan G, et al. Emoe: Modality-specific enhanced dynamic emotion experts[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 14314-14324.