DISENTANGLED REPRESENTATION LEARNING WITH DUAL-BRANCH ENHANCEMENT AND ADAPTIVE CROSS-MODAL INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sentiment Analysis (MSA) aims to capture human sentiment by jointly leveraging heterogeneous information from text, speech, and vision. While existing feature disentanglement methods partially separate modality-specific and modality-invariant representations, there remain two key limitations: (i) the disentangled representations lack targeted enhancement. Residual redundant temporal dependencies and intra-modal noise impair the discriminative capability, while insufficient cross-modal semantic interaction result in incomplete shared semantics, and (ii) the interaction features generated by cross-modal transformers are typically treated equally, which may amplify redundant and noisy dependencies while diluting the contribution of critical sentiment cues. To address these issues, we propose a disentangled representation learning framework with dual-branch enhancement and adaptive cross-modal interaction. Specifically, we first design a Dual-Branch Masked Representation Enhancement (DBMRE) module, which employs intra-modal masked attention and inter-modal masked attention to enhance modality-specific and modality-invariant representations, respectively. We then propose an Adaptive Cross-modal Interaction Gating (ACIG) mechanism, which dynamically estimates channel-wise importance weights based on the global semantic context of each interaction branch, selectively enhancing sentiment-relevant cross-modal dependencies while suppressing redundant interactions. Furthermore, we introduce multi-level knowledge distillation and dictionary-guided contrastive learning to provide auxiliary supervision at different representation levels. Extensive experiments demonstrate that our proposed framework achieves competitive performance. Furthermore, comprehensive ablation studies and training dynamics analyses validate the synergistic effectiveness of DBMRE and ACIG in enhancing both discriminative representation learning and model convergence.

Keywords: Feature Disentanglement; Dual-Branch Masked Representation Enhancement; Adaptive Cross-modal Interaction Gating; Knowledge Distillation; Contrastive Learning

1 INTRODUCTION

With the rapid proliferation of social media and intelligent human–computer interaction, multimodal data comprising textual, acoustic, and visual information have become increasingly prevalent. Consequently, Multimodal Sentiment Analysis (MSA) has emerged as a cornerstone in affective computing, with broad applications in conversational and recommender systems [Al-Dujaili et al., 2024]. However, real-world multimodal data often suffers from modality-specific noise and information imbalance [Zhuang et al., 2026]. Furthermore, modality heterogeneity, task-irrelevant interference, and inter-modal conflicts remain primary bottlenecks that limit the quality of multimodal sentiment representations [Liu and Wang, 2026; Li and Li, 2025].

To address these challenges, recent studies have adopted Transformer-based architectures to model cross-modal relationships [Wu et al., 2024], while exploring structured representations through feature disentanglement [Jian et al., 2026]. Despite these advancements, two inherent limitations remain, as illustrated in Figure. 1.

First, while feature disentanglement separates modality-specific and modality-invariant information, the resulting representation spaces lack targeted enhancement. Modality-specific representations are often impaired by task-irrelevant noise, whereas modality-invariant representations suffer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

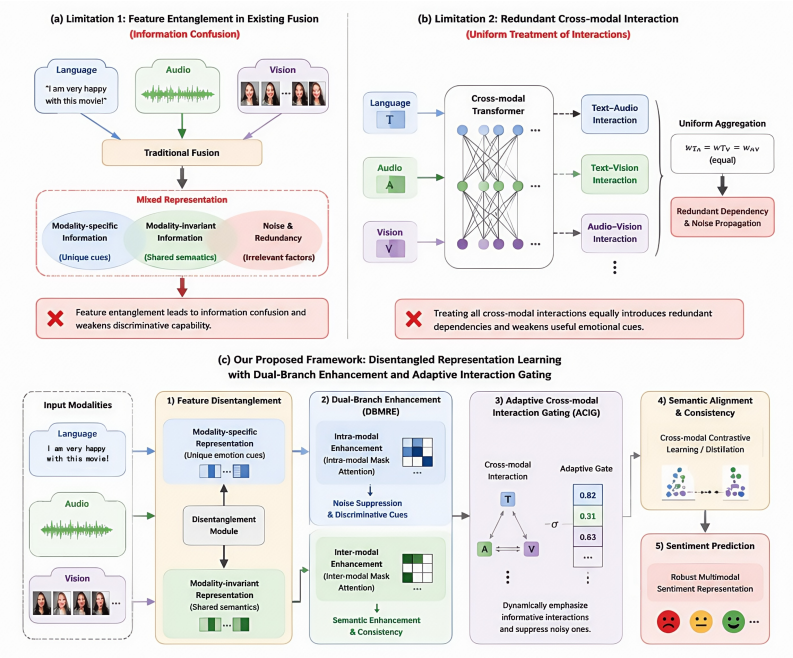

Figure 1: Motivation and overall solution of the proposed method.

from insufficient cross-modal semantic interaction and incomplete shared semantics, which collectively limit the discriminative capability of the disentangled representations [Liu and Wang, 2026].

Second, conventional cross-modal interaction mechanisms treat different interaction features uniformly, leading to the propagation of redundant dependencies and the dilution of informative sentiment cues. Modality competition and distributional differences frequently cause redundant information to be over-amplified, while complementary sentiment-relevant cues are suppressed or ignored [Li and Liu, 2025; He et al., 2026; Xu et al., 2025].

To overcome these limitations, we propose a novel disentangled representation learning framework equipped with dual-branch enhancement and adaptive cross-modal interaction for MSA. Specifically, we first design a Dual-Branch Masked Representation Enhancement (DBMRE) module following feature disentanglement. DBMRE employs an intra-modal mask branch to suppress redundant intra-modal noise and an inter-modal mask branch to strengthen cross-modal semantic consistency, thereby jointly optimizing both representation types. Furthermore, we introduce an Adaptive Cross-modal Interaction Gating (ACIG) mechanism after the multimodal Transformer. ACIG dynamically estimates channel-wise importance weights based on the global semantic context of each interaction branch, selectively amplifying sentiment-relevant dependencies while suppressing redundant interactions. Additionally, multi-level knowledge distillation and dictionary-guided contrastive learning are incorporated to provide robust auxiliary supervision.

The main contributions of this work are summarized as follows: (1) We propose a Dual-Branch Masked Representation Enhancement (DBMRE) module that employs dedicated intra-modal and inter-modal attention branches to jointly refine modality-specific and modality-invariant representations, effectively addressing the limitation of insufficient representation quality after disentanglement. (2) We propose an Adaptive Cross-modal Interaction Gating (ACIG) mechanism to dynamically regulate Transformer-generated interaction features, suppressing redundant cross-modal dependencies while selectively highlighting critical sentiment cues. (3) Extensive experiments on benchmark datasets demonstrate that our framework achieves competitive performance. Comprehensive ablation studies and training dynamics analyses validate the synergistic effectiveness of DBMRE and ACIG. The project page is available at https://github.com/lin20021113/DBMRE-ACIG, which provides the source code and detailed implementation for reproducibility.

2 Related Work

Multimodal Sentiment Analysis. Early MSA methods focus on feature-level fusion (TFN, LMF, MFN) and later Transformer-based cross-modal interaction (MulT). Recent works explore robust

and adaptive paradigms, but still lack fine-grained optimization for disentangled representation spaces, which remains a key challenge for fine-grained sentiment analysis.

Feature Disentanglement. MISA separates modality-specific and modality-invariant subspaces, and ConFEDE integrates contrastive learning. However, existing methods emphasize separation while neglecting targeted enhancement of the resulting representations.

Adaptive Cross-modal Interaction. Methods such as MFON, D2R, and KuDA adjust modality contributions, while Semi-IIN uses masked attention for interaction selection. Unlike these, our ACIG operates on interaction features after disentanglement and DBMRE, performing channel-wise adaptive gating. A comprehensive review of related work is provided in Appendix A.

3 METHODOLOGY

3.1 OVERALL FRAMEWORK

Built upon a robust feature disentanglement and adaptive interaction paradigm, the proposed model consists of six sequential stages: feature extraction and disentanglement, Dual-Branch Masked Representation Enhancement (DBMRE), coarse-grained graph distillation, Adaptive Cross-modal Interaction Gating (ACIG), fine-grained dictionary matching, and multimodal fusion for sentiment prediction, which together yield highly discriminative and robust sentiment representations.

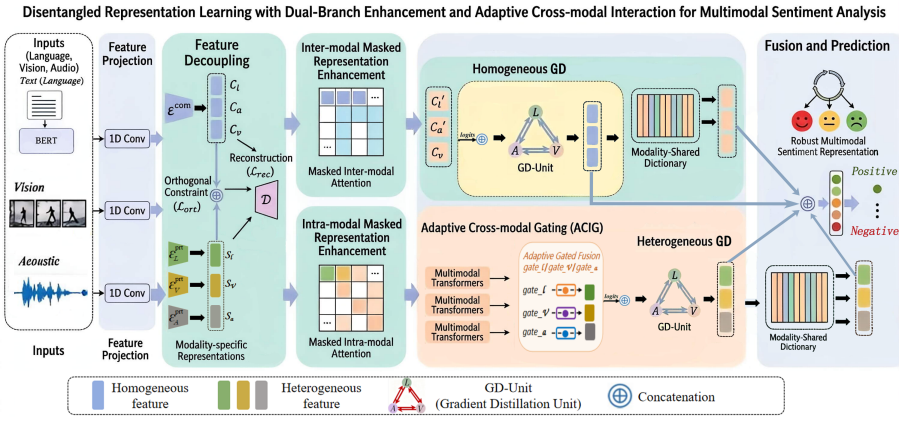

Figure 2: Overall architecture of the proposed multimodal sentiment analysis framework.

Given an input video utterance, the model processes text, speech, and visual modalities. The textual modality is encoded by a pre-trained BERT, while the speech and visual modalities use handcrafted low-level features. Three independent 1D convolutional layers project these features into a shared latent space. Then, a feature disentanglement module decomposes each modality into modality-specific and modality-invariant representations.

To enhance the quality of disentangled representations, we introduce the DBMRE module, which applies intra-modal and inter-modal masked attention to refine modality-specific and modality-invariant representations, respectively. The enhanced representations are then fed into a cross-modal Transformer with ACIG, which dynamically selects informative interaction features. Finally, a shared dictionary matching module aligns and reconstructs the representations, and the fused features are used for sentiment regression. The entire framework is trained end-to-end with multiple complementary objectives.

3.2 FEATURE DISENTANGLEMENT

3.2.1 MULTIMODAL FEATURE REPRESENTATION

Formally, given a multimodal dataset $\mathcal{D} = \{(X_l^{(i)}, X_a^{(i)}, X_v^{(i)}, y^{(i)})\}_{i=1}^{N}$, where X_l, X_a, X_v denote the textual, acoustic, and visual modalities respectively, our objective is to learn a model f_θ that minimizes the following empirical risk:

$$\min_{\theta} \mathbb{E}_{(X,y)\sim\mathcal{D}}[\ell(f_\theta(X), y)] + \Omega(\theta)$$

where $\ell(\cdot)$ denotes the task loss and $\Omega(\theta)$ denotes the regularization term.

Multimodal sentiment analysis aims to jointly exploit complementary sentiment cues from textual, acoustic, and visual modalities. However, due to substantial discrepancies in feature distributions, semantic granularity, and temporal dynamics across modalities, directly performing multimodal fusion often leads to sub-optimal performance. Therefore, prior to representation learning, it is imperative to project heterogeneous modality features into a shared latent space while preserving their intrinsic semantic characteristics for subsequent disentanglement and sentiment prediction.

Formally, given an input utterance, a multimodal sample can be represented as $X = \{X_l, X_a, X_v\}$, where X_l, X_a, X_v denote the textual, acoustic, and visual modalities, respectively. For the textual modality, we employ a pre-trained BERT encoder to extract rich contextual semantic information, whereas the acoustic and visual modalities are represented using low-level feature descriptors extracted from raw video streams. To alleviate cross-modal dimensional discrepancies, three independent one-dimensional convolutional projection layers are employed to map features from different modalities into a shared embedding space, formulated as shown in Equation (1).

$$F_m = \mathrm{Conv}_m(X_m), \quad m \in \{l, a, v\} \tag{1}$$

where $\mathrm{Conv}_m(\cdot)$ denotes the feature projection layer of the corresponding modality, and $F_m \in \mathbb{R}^{T_m \times d}$ denotes the projected feature representation with a unified embedding dimension of d. Through this projection operation, heterogeneous inputs are transformed into consistent dimensions, providing a unified foundation for subsequent disentangled representation learning.

3.2.2 Disentangled Feature Learning

Although unified feature projection mitigates distribution gaps, the projected representations still entangle modality-specific sentiment traits with modality-invariant semantic information. Directly modeling these entangled representations in a shared space often induces severe semantic interference: modality-dependent noise can contaminate shared semantics, while shared representations may suppress modality-specific cues. Consequently, the overall discriminative capacity is severely compromised, which further degrades the quality of downstream multimodal sentiment representations and prediction performance.

To address this issue, we adopt a feature disentanglement strategy that explicitly separates modality-specific and modality-invariant representations via two parallel encoding branches. Specifically, each projected feature is simultaneously fed into a modality-specific encoder and a modality-invariant encoder, as shown in Equation (2).

$$S_m = E_s(F_m), \quad C_m = E_c(F_m) \tag{2}$$

where $E_s(\cdot)$ and $E_c(\cdot)$ denote the modality-specific and modality-invariant encoders, respectively, and $m \in \{l, a, v\}$ denotes the corresponding output representation.

The modality-specific branch preserves unique affective characteristics that cannot be inferred from other modalities, whereas the modality-invariant branch captures core semantic information shared across modalities. By explicitly decoupling these two complementary components, our framework effectively reduces inter-modal interference and constructs a more structured representation space for downstream multimodal interaction and sentiment prediction.

To prevent the two branches from learning partially overlapping information during optimization, we further introduce a cosine similarity-based decorrelation constraint, as formulated in Equation (3), which encourages the two branches to capture complementary information.To measure the correlation between the modality-specific and modality-invariant representations, the cosine similarity is computed as: where the cosine similarity is computed as $\cos(S_m, C_m) = \frac{S_m^\top C_m}{\|S_m\|_2 \|C_m\|_2}$. Then, the orthogonal constraint is defined as:

$$\mathcal{L}_{ort} = \sum_{m \in \{l,a,v\}} \max(0, \cos(\mathbf{S}_m, \mathbf{C}_m)) \tag{3}$$

Minimizing $\mathcal{L}_{ort}$ reduces the mutual correlation between the two branches, encouraging them to capture strictly complementary information and thereby enhancing the effectiveness of feature disentanglement. To ensure that the reconstructed representation preserves original semantics, the projected features are used as supervision signals via a reconstruction objective, which is formally defined in Equation (5) and explicitly constrains the reconstructed features to remain close to their original representations.

$$\mathcal{L}_{rec} = \sum_m \|\hat{Z}_m - Z_m\|_2^2 \tag{5}$$

3.3 Dual-Branch Masked Representation Enhancement

3.3.1 Motivation

Although feature disentanglement separates modality-specific and modality-invariant representations, directly using them for downstream interaction is sub-optimal. These two representation types serve distinct semantic roles and therefore suffer from different deficiencies.

Specifically, modality-specific representations often retain intra-modal noise and redundant temporal dependencies that erode their discriminative capability in cross-modal interactions, while modality-invariant representations lack sufficient cross-modal interaction during encoding, leading to incomplete shared semantics and further limiting the overall quality of learned multimodal representations, leading to incomplete shared semantics and further limiting the overall quality of learned multimodal representations for downstream sentiment analysis.

These observations motivate differentiated enhancement: modality-specific representations should strengthen intra-modal dependencies while suppressing cross-modal interference; modality-invariant representations should promote cross-modal semantic interaction to improve shared consistency and eliminate redundant intra-modal modeling.

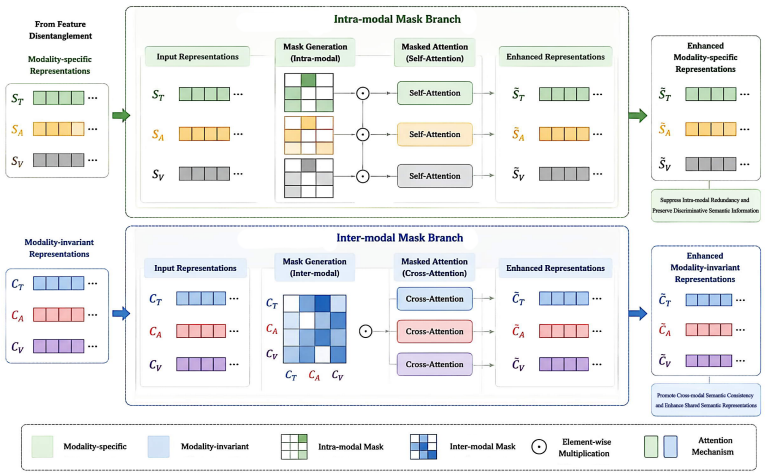

Figure 3: Illustration of the Dual-Branch Masked Representation Enhancement (DBMRE) mechanism.

Building upon this analysis, we propose the Dual-Branch Masked Representation Enhancement (DBMRE) module, as illustrated in Figure 3. DBMRE employs two complementary branches: the intra-modal branch restricts attention within each modality to suppress redundant dependencies and enhance modality-specific sentiment cues, while the inter-modal branch performs masked attention exclusively across modalities to promote shared semantic alignment and complementary learning. The two branches are integrated through residual fusion to obtain enhanced disentangled representations, providing a stronger foundation for downstream cross-modal interaction and multimodal fusion, which ultimately benefits sentiment prediction performance.

3.3.2 Intra-modal Representation Enhancement

The primary objective of the modality-specific branch is to preserve unique sentiment characteristics within each modality, avoiding unnecessary pollution from heterogeneous sources. Conventional self-attention mechanisms compute global dependencies indiscriminately across all tokens, which risks introducing cross-modal interference and degrading modality-specific discriminability.

To address this, we propose an intra-modal representation enhancement strategy that restricts attention computation strictly to intra-modal boundaries. Formally, the modality-specific representations from the three modalities are first concatenated along the temporal dimension to form a unified sequence: $S = [S_l; S_v; S_a]$, where $[\cdot;\cdot]$ denotes temporal concatenation.

Based on pre-defined modality indices, we construct a binary intra-modal attention mask that ensures each token only attends to tokens within the same modality, thereby preventing cross-modal interference. The mask is formally defined in Equation (10).

$$M_{ij}^{\text{intra}} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ belong to the same modality,} \\ -\infty, & \text{otherwise,} \end{cases} \tag{10}$$

where M_{ij}^{intra} determines whether the query token i can attend to the key token j. This mask restricts attention paths entirely within individual modalities while completely blocking cross-modal connections, and the output is then fused with the original features via residual learning, as detailed in Appendix B. This design effectively preserves the discriminative information of each modality while suppressing irrelevant cross-modal noise.

3.3.3 Inter-modal Representation Enhancement

Conversely, modality-invariant representations aim to capture shared semantics. Although the shared encoder maps different modalities into a common latent space, independent optimization leaves potential cross-modal correspondences underexploited. While applying unrestricted self-attention to concatenated sequences could model these interactions, it inevitably results in redundant re-learning of intra-modal contexts already captured by the specific branch.

To resolve this, we propose an inter-modal representation enhancement strategy that restricts attention computations exclusively across heterogeneous modalities. Specifically, the modality-invariant representations are concatenated into a unified sequence: $C = [C_l; C_v; C_a]$, where C_l, C_v, C_a denote the modality-invariant representations of the textual, visual, and speech modalities, respectively. Based on modality identities, we construct an inter-modal attention mask that only allows each token to attend to tokens from other modalities, thereby promoting cross-modal semantic interaction while suppressing intra-modal redundancy. The mask is formally defined in Equation (13), which is essential for capturing cross-modal semantic dependencies and enhancing shared representation learning.

$$M_{ij}^{\text{inter}} = \begin{cases} -\infty, & \text{if } i \text{ and } j \text{ belong to the same modality,} \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where i and j denote two positions in the multimodal sequence.Based on the mask, the attention weight A_{ij} and the scaled dot-product e_{ij} are computed as:

$$A_{ij} = \frac{\exp(e_{ij} + M_{ij})}{\sum_k \exp(e_{ik} + M_{ik})}, \quad e_{ij} = \frac{q_i^\top k_j}{\sqrt{d}}$$

where q_i and k_j denote the query and key vectors, and d is their dimension. This mask zeroes out intra-modal attention weights post-softmax, isolating the attention capacity entirely for cross-modal semantic interaction. Under this constraint, the masked attention is formulated as Equation (14).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M^{\text{inter}}\right)V \tag{14}$$

The output is partitioned and preserved via residual learning, as detailed in Appendix B.

3.3.4 Complementary Representation Enhancement

In summary, DBMRE addresses distinct post-disentanglement deficiencies through two coordinated branches. The intra-modal branch acts as a representation purifier, while the inter-modal branch serves as a semantic binder. Their collaborative operation ensures that modality-specific and modality-invariant representations are optimally refined prior to downstream cross-modal interactions, providing a solid foundation for subsequent multimodal fusion and sentiment prediction.

However, while DBMRE successfully optimizes individual representation spaces, subsequent cross-modal interactions still lack the ability to dynamically weigh diverse interaction pathways. To bridge this gap, we introduce the Adaptive Cross-modal Interaction Gating (ACIG) mechanism, which selectively reweights cross-modal interaction features prior to memory aggregation.

3.4 Adaptive Cross-modal Interaction Gating

Although DBMRE refines disentangled representations, the subsequent cross-modal Transformer still constructs multiple parallel interaction branches with varying sentiment relevance. Uniform aggregation treats all branches equally, potentially amplifying noisy interactions and suppressing informative sentiment cues. Therefore, dynamically evaluating each branch's contribution prior to memory aggregation is crucial.

To address this, we propose ACIG. It first aggregates the global semantic context of each interaction branch, then learns channel-wise importance weights via a parameterized gating function.

These weights are used to adaptively reweight the interaction features before they are fed into the memory Transformer. Consequently, informative responses are amplified while redundant or noisy cues are suppressed. Specifically, ACIG comprises four stages: interaction feature construction, global context aggregation, adaptive gating generation, and interaction feature reweighting.

The interaction features are constructed as detailed in Appendix B. To mitigate this issue, we employ global average pooling to summarize the overarching semantic context of each interaction branch, as shown in Equation (19).

$$\bar{H}_l = \frac{1}{T}\sum_{t=1}^{T} H_l^{(t)} \tag{19}$$

where T denotes the sequence length.Global average pooling effectively distills the holistic semantic information within each interaction branch into a compact feature representation, providing a robust foundation for subsequent gating weight computation.

Building upon the aggregated interaction context, we introduce a parameterized gating function to estimate the importance coefficients across different interaction channels. Specifically, the adaptive gating weight computation is formulated as shown in Equation (20).

$$G_l = \sigma(W_2\delta(W_1\bar{H}_l)) \tag{20}$$

where W_1 and W_2 denote the two fully connected layers, respectively; δ denotes the ReLU activation function; and σ denotes the Sigmoid activation function.

Constrained by the Sigmoid function, the gating weights are restricted to the range $[0, 1]$, empowering the model to adaptively quantify the significance of each interaction channel. Subsequently, these estimated gating weights are applied to the interaction features via channel-wise multiplication for feature recalibration, as described in Equation (21).

$$\tilde{H}_l = G_l \odot H_l \tag{21}$$

where $\odot$ denotes the element-wise multiplication operation.The gated features are then fed into the memory Transformer for higher-order temporal modeling, as detailed in Appendix B.

Crucially, ACIG and the aforementioned DBMRE module complement each other from distinct perspectives. Specifically, DBMRE primarily elevates multimodal representation quality through structured attention constraints, whereas ACIG optimizes the cross-modal interaction process via adaptive feature reweighting. Together, these two modules form a progressive representation learning framework: DBMRE performs feature-level optimization by designing tailored attention constraints for diverse semantic representations, while ACIG executes interaction-level selection and enhancement by dynamically adjusting cross-modal interaction weights. Their synergistic operation equips the model with more discriminative multimodal sentiment representations, laying a solid feature foundation for subsequent sentiment prediction.

3.5 Multi-level Knowledge Distillation

To provide auxiliary supervision for intermediate representations, we introduce homogeneous and heterogeneous distillation branches, together with a dictionary-guided contrastive learning objective. The overall knowledge distillation loss is $\mathcal{L}_{KD} = \mathcal{L}_{\text{homo}} + \mathcal{L}_{\text{hetero}}$. Detailed formulations of each distillation objective are provided in Appendix B.

3.6 Overall Optimization Objective

The proposed framework is optimized end-to-end by jointly integrating the primary sentiment prediction objective, feature disentanglement objective, multi-level knowledge distillation objective, and dictionary-guided semantic learning objective, which together form a unified multi-task optimization framework. The overall training objective is formally defined in Equation (37):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{FD} + \lambda_2\mathcal{L}_{KD} + \lambda_3\mathcal{L}_{DM} \tag{37}$$

where $\mathcal{L}_{\text{task}}$, $\mathcal{L}_{FD}$, $\mathcal{L}_{KD}$, and $\mathcal{L}_{DM}$ denote the primary prediction, feature disentanglement (Section 3.2), multi-level knowledge distillation (Section 3.5), and dictionary-guided semantic learning objectives, respectively. The hyperparameters λ_1, λ_2, and λ_3 control their relative contributions. By jointly optimizing these complementary objectives, the framework achieves effective feature disentanglement, enhanced multimodal representation learning, adaptive knowledge transfer, and accurate sentiment prediction.

4 EXPERIMENTS AND ANALYSIS

4.1 EXPERIMENTAL SETUP

We evaluate our framework on two widely adopted benchmark datasets for multimodal sentiment analysis: CMU-MOSI [30] and CMU-MOSEI [31]. Both datasets provide continuous sentiment intensity annotations and serve as standard testbeds for multimodal affective computing. For evaluation, we adopt standard metrics including MAE, Corr, Acc-2, F1, and Acc-7. Detailed dataset statistics, formal metric definitions, class mapping rules, implementation details, network hyperparameters, and loss weights are summarized in Appendix C.

4.2 COMPARISON WITH EXISTING METHODS

We compare our framework with representative methods on CMU-MOSI and CMU-MOSEI, including classical fusion methods (TFN, LMF, MFN, MulT), feature disentanglement methods (MISA, Self-MM), and recent attention- and gating-enhanced methods (MTAG, GraphCAGE). We also reproduce DHMD as a strong baseline and incrementally add DBMRE and ACIG to build our full model, which allows us to isolate the contribution of each proposed module and verify their effectiveness on a strong recent baseline.

Table 1: Comparison results with existing methods on CMU-MOSI and CMU-MOSEI. The best results are highlighted in bold, while the second-best results are underlined.

Method	**CMU-MOSI**					**CMU-MOSEI**				
	Acc-2	**F1**	**Acc-7**	**MAE**	**Corr**	**Acc-2**	**F1**	**Acc-7**	**MAE**	**Corr**
TFN [Zadeh et al., 2017]	78.3	78.2	33.7	0.925	0.662	81.0	81.1	52.2	0.570	0.716
LMF [Liu et al., 2018]	77.5	77.3	32.7	0.931	0.670	81.3	81.6	52.0	0.568	0.727
MFN [Zadeh et al., 2018a]	77.9	77.8	35.8	0.927	0.670	78.9	79.0	51.3	0.573	0.718
MulT [Tsai et al., 2019]	80.6	79.3	35.5	0.972	0.681	81.4	81.7	49.0	0.630	0.664
MISA [Hazarika et al., 2020]	81.8	81.7	43.5	0.752	0.784	81.6	82.0	52.2	0.550	0.758
Self-MM [Yu et al., 2021]	82.7	82.6	45.8	0.731	0.785	82.6	82.8	50.6	0.547	0.763
MMIM [Han et al., 2021]	83.0	82.9	45.0	0.738	0.781	81.9	82.3	53.1	0.547	0.752
ConFEDE [Yang et al., 2023]	84.4	84.4	46.3	0.741	0.783	82.8	83.1	53.1	0.538	0.771
ALMT [Zhang et al., 2023]	83.4	83.2	46.8	0.725	0.787	83.3	82.9	53.0	0.543	0.765
TETFN [Wang et al., 2023]	83.2	83.1	45.8	0.708	0.798	84.1	84.4	53.9	0.537	0.767
ULMD [Zhu, L. et al., 2025]	83.1	82.9	47.8	**0.700**	0.799	82.6	83.0	53.8	**0.531**	0.771
DEVA [Wu et al., 2025]	84.4	84.5	46.3	0.730	0.787	83.3	82.9	52.3	0.541	0.769
MTAG [Yang et al., 2021]	80.5	80.4	31.9	0.941	0.692	79.1	75.9	48.2	0.645	0.614
GraphCAGE [Wu et al., 2021]	82.1	82.1	35.4	0.933	0.684	81.7	81.8	48.9	0.609	0.670
DHMD [Li et al., 2026]	83.69	83.66	46.16	0.716	0.782	84.06	84.12	51.35	0.614	0.762
Ours	**84.86**	**84.92**	**47.90**	0.711	**0.803**	**85.17**	**85.18**	52.44	0.592	**0.780**

Note: ① The left half reports the results on CMU-MOSI, while the right half reports the results on CMU-MOSEI. ② Except for DHMD and our method, the results of the other baselines are taken from the corresponding results publicly reported in their original publications. ③ The DHMD results are independently reproduced in this study based on its publicly available implementation, while the proposed method (Ours) is obtained by incorporating DBMRE and ACIG into the reproduced DHMD framework. ④ Different studies may adopt different feature extraction methods, modality alignment strategies, and evaluation settings; therefore, the comparisons are based on the corresponding experimental settings reported in the respective publications.

As shown in Table 1, our method achieves superior performance across multiple evaluation metrics on both datasets. Compared with classical fusion methods, it demonstrates substantial gains, validating the effectiveness of the progressive "disentanglement to dual-branch enhancement to adaptive interaction selection" paradigm. Compared with disentanglement-based methods, DBMRE and ACIG further improve representation quality and suppress redundant interactions. Due to space limitations, we provide detailed baseline descriptions and a comprehensive result analysis in Appendix D to further support these observations and validate the effectiveness of our approach.

4.3 ABLATION STUDIES

To systematically validate the contribution and necessity of each component within our framework, we conduct exhaustive ablation studies primarily on the CMU-MOSI dataset. We adopt the reproduced DHMD as our reference baseline, which incorporates feature disentanglement, knowledge distillation, and dictionary-guided contrastive learning, but omits the proposed Dual-Branch Masked Representation Enhancement (DBMRE) and Adaptive Cross-modal Interaction Gating

(ACIG) modules. By incrementally integrating our proposed components and evaluating alternative design strategies, we isolate their individual and synergistic effects. The quantitative results are summarized in Tables 2 and 3.

Table 2: Ablation study results on CMU-MOSI (mean ± standard deviation)

Model	DBMRE	ACIG	Acc-2(%)	F1(%)	Acc-7(%)	MAE	Corr
DHMD	×	×	83.69 ± 0.12	83.66 ± 0.12	46.16 ± 0.65	0.716 ± 0.0082	0.782 ± 0.0032
DHMD+DBMRE	✓	×	85.06 ± 0.46	85.06 ± 0.42	46.58 ± 0.08	0.713 ± 0.0066	0.797 ± 0.0025
DHMD+ACIG	×	✓	83.54 ± 0.30	83.54 ± 0.34	46.28 ± 0.08	0.737 ± 0.0014	0.791 ± 0.0030
Ours	✓	✓	84.86 ± 0.31	84.92 ± 0.38	47.90 ± 0.30	0.711 ± 0.0015	0.803 ± 0.0013

Table 3: Ablation results of different DBMRE and ACIG design strategies (mean ± standard deviation)

Model	Acc-2 (%)	F1 (%)	Acc-7 (%)	MAE
+ Dual-Branch Masking + Linear Gating	83.46 ± 0.08	83.48 ± 0.07	46.06 ± 0.30	0.729 ± 0.0017
+ GCN-based Cross-modal Message Passing + Adaptive Gating	83.84 ± 0.15	83.85 ± 0.15	45.48 ± 0.29	0.727 ± 0.0071
+ Random Feature Masking + Adaptive Gating	83.46 ± 0.08	83.48 ± 0.10	45.40 ± 0.36	0.728 ± 0.004
+ Single-Branch Masking (S Branch Only) + Adaptive Gating	84.60 ± 0.30	84.62 ± 0.33	47.38 ± 0.28	0.718 ± 0.0029
+ Single-Branch Masking (C Branch Only) + Adaptive Gating	83.76 ± 0.56	83.78 ± 0.68	45.92 ± 0.58	0.724 ± 0.0185

Table 2 shows that adding DBMRE alone improves Acc-2 by 1.37 percentage points, while adding ACIG alone degrades MAE, confirming that interaction selection requires high-quality representations. The full model achieves the best Acc-7 and lowest MAE. Table 3 further validates that our dual-branch masking and adaptive gating outperform linear gating, GCN-based message passing, and random masking. Detailed ablation analysis is provided in Appendix D.

4.4 Discussion

Due to space limitations, the detailed training dynamics (Figs. 4–6), computational complexity analysis, and the extended discussion on limitations and future work are provided in Appendix E, Appendix F, and Appendix G, respectively, to ensure the completeness of our presentation.

In brief, our analysis shows that the adaptive gating variant exhibits smoother optimization trajectories and superior convergence stability, validating its effectiveness in regularizing the interaction space. Furthermore, as shown in Table 3, combining our dual-branch masking with adaptive gating (the full model) achieves the best Acc-7 and lowest MAE, outperforming variants that use linear gating, GCN-based message passing, or random masking. This confirms that the proposed DBMRE and ACIG modules work synergistically: DBMRE purifies the representations, while ACIG selectively amplifies informative cross-modal interactions. Meanwhile, modality-specific cues are found to play a decisive role in fine-grained sentiment prediction, although modality-invariant representations provide necessary complementary semantics. Despite these promising results, our method still has some limitations. The dual-branch masking and adaptive gating mechanisms introduce additional hyperparameters, and the computational cost may increase when scaling to longer sequences. In future work, we plan to explore more lightweight gating designs and extend our framework to more challenging scenarios such as missing modalities and cross-domain sentiment analysis.

5 Conclusion

In this work, we proposed a disentangled representation learning framework for multimodal sentiment analysis, centered on DBMRE and ACIG. DBMRE enhances intra-modal discriminative semantics and inter-modal consistency via dedicated masking branches, while ACIG dynamically refines cross-modal interactions via global context-aware channel-wise weighting. Extensive experiments and ablation studies validate their synergistic contributions and state-of-the-art performance. Future work will explore lightweight masking and gating mechanisms and extend the framework to multilingual and cross-domain sentiment understanding. Detailed discussions on limitations and potential extensions are provided in Appendix G.

REFERENCES

A. A. Al-Dujaili, et al. Progress, achievements, and challenges in multimodal sentiment analysis using deep learning. Applied Soft Computing, 152:111206, 2024.

Y. Chen, K. Li, W. Mai, et al. D2R: Dual-branch dynamic routing network for multimodal sentiment detection. In Proceedings of EMNLP, pp. 3536–3547, 2024.

X. Feng, Y. Lin, L. He, et al. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In Findings of EMNLP, pp. 14755–14766, 2024.

Z. Guo, T. Jin, and Z. Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Proceedings of ACL, pp. 1726–1736, 2024.

W. Han, H. Chen, and S. Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Proceedings of EMNLP, pp. 9180–9192, 2021.

D. Hazarika, R. Zimmermann, and S. Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In Proceedings of ACM MM, pp. 1122–1131, 2020.

K. He, B. Chen, Y. Ding, et al. PaSE: Prototype-aligned calibration and Shapley-based equilibrium for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, 2026.

Z. Jian, W. Lv, Y. Chen, et al. MDF: A modality-aware disentanglement and fusion framework for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 31292–31300, 2026.

M. Li, D. Yang, X. Zhao, et al. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In Proceedings of CVPR, pp. 12458–12468, 2024.

T. Li and D. Liu. MPID: A modality-preserving and interaction-driven fusion network for multimodal sentiment analysis. In Proceedings of COLING, 2025.

X. Li, X. Cheng, D. Miao, et al. TF-Mamba: Text-enhanced fusion Mamba with missing modalities for robust multimodal sentiment analysis. In Findings of EMNLP, pp. 11252–11267, 2025.

Y. Li, Y. Wang, and Z. Cui. Decoupled multimodal distilling for emotion recognition. In Proceedings of CVPR, pp. 6631–6640, 2023.

Y. Li, Y. Wang, Y. Ding, et al. Decoupled hierarchical distillation for multimodal emotion recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

Z. Li and L. Li. t-HNE: A text-guided hierarchical noise eliminator for multimodal sentiment analysis. In Proceedings of COLING, pp. 2834–2844, 2025.

J. Lin, Y. Wang, Y. Xu, et al. Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1411–1419, 2025.

Y. Liu and S. Wang. FINE: Factorized multimodal sentiment analysis via mutual information estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 1955–1963, 2026.

Z. Liu, Y. Shen, V. B. Lakshminarasimhan, et al. Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of ACL, pp. 2247–2256, 2018.

K. Sun and M. Tian. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In Proceedings of COLING, pp. 40–49, 2025.

Y. H. H. Tsai, S. Bai, P. P. Liang, et al. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of ACL, pp. 6558–6569, 2019.

D. Wang, X. Guo, Y. Tian, et al. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition, 136:109259, 2023.

Y. Wei, S. Yuan, R. Yang, et al. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In Proceedings of ACL, pp. 5240–5252, 2023.

J. Wu, S. Mai, and H. Hu. Graph capsule aggregation for unaligned multimodal sequences. In Proceedings of the International Conference on Multimodal Interaction, pp. 151–159, 2021.

S. Wu, X. Wang, L. Wang, et al. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1601–1609, 2025.

Z. Wu, Z. Gong, J. Koo, et al. Multimodal multi-loss fusion network for sentiment analysis. In Proceedings of NAACL, pp. 3588–3602, 2024.

Y. Xie, Z. Zhu, X. Lu, et al. InfoEnh: Towards multimodal sentiment analysis via information bottleneck filter and optimal transport alignment. In Proceedings of LREC-COLING, pp. 9073–9083, 2024.

Q. Xu, Y. Wei, C. Wu, et al. Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In Proceedings of the AAAI Conference on Artificial Intelligence, 2025.

J. Yang, Y. Wang, R. Yi, et al. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In Proceedings of NAACL, pp. 1009–1021, 2021.

J. Yang, Y. Yu, D. Niu, et al. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In Proceedings of ACL, pp. 7617–7630, 2023.

Y. Yang, X. Dong, and Y. Qiang. CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In Findings of NAACL, pp. 2099–2110, 2024.

W. Yu, H. Xu, F. Meng, et al. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Proceedings of ACL, pp. 3718–3727, 2020.

W. Yu, H. Xu, Z. Yuan, et al. Self-MM: Self-supervised multi-task learning with joint label refinement for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6219–6227, 2021.

A. Zadeh, R. Zellers, E. Pincus, et al. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In Proceedings of Interspeech, pp. 1316–1320, 2016.

A. Zadeh, M. Chen, S. Poria, et al. Tensor fusion network for multimodal sentiment analysis. In Proceedings of EMNLP, pp. 1103–1114, 2017.

A. Zadeh, P. P. Liang, N. Mazumder, et al. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5634–5641, 2018a.

A. Zadeh, P. P. Liang, S. Poria, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Proceedings of ACL, pp. 2236–2246, 2018b.

H. Zhang, Y. Wang, G. Yin, et al. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Proceedings of EMNLP, pp. 756–767, 2023.

X. Zhang, W. Wei, and S. Zou. Modal feature optimization network with prompt for multimodal sentiment analysis. In Proceedings of the 31st International Conference on Computational Linguistics, pp. 4611–4621, 2025.

M. Zhou, L. Yang, T. Wu, et al. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In Proceedings of EMNLP, pp. 11355–11365, 2025.

A. Zhu, M. Hu, X. Wang, et al. Multimodal invariant sentiment representation learning. In Findings of ACL, pp. 14743–14755, 2025a.

A. Zhu, M. Hu, X. Wang, et al. Proxy-driven robust multimodal sentiment analysis with incomplete data. In Proceedings of ACL, pp. 22123–22138, 2025b.

L. Zhu, H. Zhao, Z. Zhu, et al. Multimodal sentiment analysis with unimodal label generation and modality decomposition. Information Fusion, 116:102787, 2025.

Y. Zhuang, M. Liu, Y. Zhang, et al. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, 2026.

APPENDIX

The appendix is organized as follows:

- A Detailed Related Work
- B Detailed Methodology and Derivations
- C Datasets and Implementation Details
- D Additional Experimental Results
- E Training Dynamics
- F Computational Complexity Analysis
- G Discussion and Limitations

A DETAILED RELATED WORK

A.1 MULTIMODAL SENTIMENT ANALYSIS

Multimodal Sentiment Analysis (MSA) aims to capture human sentiment by jointly leveraging heterogeneous information from text, speech, and vision. Compared with unimodal approaches, multimodal learning effectively exploits cross-modal complementary cues, leading to more comprehensive sentiment representations. Consequently, MSA has become a cornerstone in affective computing and multimodal machine learning [Al-Dujaili et al., 2024].

Early studies primarily focused on fusion strategies, such as feature-level and decision-level fusion. Representative models include Tensor Fusion Network (TFN) [Zadeh et al., 2017], Low-rank Multimodal Fusion (LMF) [Liu et al., 2018], and Memory Fusion Network (MFN) [Zadeh et al., 2018a], which capture high-order interactions and temporal dynamics. However, simple feature concatenation or fixed-weight aggregation often introduces inter-modal noise and fails to model complex semantic relationships. With the advent of deep learning, Transformer-based architectures have dominated MSA by employing self-attention mechanisms to learn dynamic cross-modal dependencies [Wu et al., 2024]. Despite their success, standard Transformers typically treat multimodal interactions indiscriminately, leading to the propagation of redundant dependencies. Furthermore, how to structure and purify representations before cross-modal interaction remains a critical challenge. Recent advances have shifted toward robust, adaptive, and missing-robust paradigms, employing prompt learning [Guo et al., 2024], proxy modalities [Zhu et al., 2025b], and efficient sequence modeling [Li, X. et al., 2025]. Nevertheless, these methods still lack fine-grained optimization for disentangled representation spaces.

A.2 FEATURE DISENTANGLEMENT-BASED MULTIMODAL REPRESENTATION LEARNING

Since raw multimodal data contain both modality-specific and modality-shared information, feature disentanglement has emerged as a prominent technique to mitigate cross-modal semantic interference. Pioneering works like MISA [Hazarika et al., 2020] map representations into distinct subspaces, where modality-specific spaces preserve unique affective traits (e.g., acoustic intonation or visual expressions), and modality-invariant spaces capture shared core semantics.

Subsequent studies have integrated contrastive learning to further sharpen the boundaries between common and distinctive features. For instance, ConFEDE [Yang et al., 2023] and multi-view calibration networks [Wei et al., 2023] explicitly decouple representations to alleviate representation shifts and annotation inconsistencies. Despite these advancements, current disentanglement methods suffer from an inherent limitation: they focus heavily on the separation process itself while neglecting the targeted enhancement of the resulting representations. Although information bottleneck techniques like InfoEnh attempt to filter noise [Xie et al., 2024], modality-specific spaces often retain residual intra-modal noise, while invariant spaces suffer from insufficient cross-modal semantic interaction. Recent efforts like MISR [Zhu et al., 2025a] and SFTTR [Sun and Tian, 2025] address stability or text-dissimilar alignment, but they still lack dedicated mechanisms to refine disentangled spaces according to their unique semantic properties. To bridge this gap, our work introduces the DBMRE module to perform targeted dual-branch enhancement on both disentangled representations.

A.3 CROSS-MODAL INTERACTION AND ADAPTIVE FUSION METHODS

Cross-modal interaction is vital for establishing effective information pathways among modalities. While cross-modal Transformers (e.g., MulT [Tsai et al., 2019]) excel at modeling unaligned

sequences, they often allow unrestricted information exchange, which can amplify irrelevant noise and dilute crucial sentiment cues.

To tackle this, recent research has explored adaptive and dynamic fusion mechanisms. Methods such as MFON [Zhang, X. et al., 2025], D2R [Chen et al., 2024], and KuDA [Feng et al., 2024] utilize modality prompts, dynamic routing, or attention to adjust modality contributions. More recent frameworks, such as DPDF-LQ [Zhou et al., 2025] and Semi-IIN [Lin et al., 2025], incorporate global-local pathways and masked attention for fine-grained selection. However, these existing approaches typically perform dynamic weighting at the coarse modality level or directly select raw interaction features, failing to bridge the gap between disentangled representation spaces and interaction branches. Unlike prior arts, our ACIG mechanism operates directly on the interaction features generated after feature disentanglement and DBMRE, dynamically estimating channel-wise importance weights to suppress redundant dependencies while preserving core sentiment cues.

A.4 Multimodal Knowledge Distillation and Semantic Alignment Learning

Knowledge distillation (KD) and contrastive learning have been widely adopted to boost cross-modal consistency and representation robustness. In MSA, methods like DMD [Li et al., 2023] and CorrKD [Li et al., 2024] leverage heterogeneous knowledge transfer and sample-level contrastive learning to enhance cross-modal alignment. Concurrently, contrastive frameworks like CLGSI [Yang et al., 2024] refine sentiment representations by pulling consistent semantics closer.

Despite these advancements, existing KD methods primarily focus on single-level transfer or rigid response matching, leaving multi-level supervision across disentangled spaces and interaction stages underexplored. Furthermore, integrating auxiliary supervision to jointly optimize feature disentanglement, representation enhancement, and cross-modal interaction remains an open challenge. To address this, our framework incorporates multi-level knowledge distillation and dictionary-guided contrastive learning, providing robust hierarchical supervision throughout the entire progressive learning pipeline.

A.5 Summary

In summary, while prior research has made notable progress in multimodal fusion, feature disentanglement, and adaptive interaction, three primary limitations persist: (1) Existing disentanglement methods overlook the necessity of targeted enhancement for separate representation spaces, leaving residual noise and incomplete shared semantics. (2) Conventional cross-modal interaction mechanisms treat interactions uniformly, failing to dynamically suppress redundant dependencies and leading to the dilution of critical sentiment cues. (3) Current auxiliary supervision strategies lack a hierarchical, multi-level design to jointly optimize feature disentanglement and cross-modal interaction. To overcome these limitations, we propose a unified framework comprising the Dual-Branch Masked Representation Enhancement (DBMRE) module, the Adaptive Cross-modal Interaction Gating (ACIG) mechanism, and multi-level knowledge distillation and contrastive learning, achieving robust and discriminative multimodal sentiment analysis.

B Detailed Methodology and Derivations

B.1 Feature Disentanglement Details

To prevent semantic information loss during disentanglement, we introduce a reconstruction branch to ensure the fidelity of the original representations. Specifically, the modality-specific and modality-invariant representations are concatenated and fed into a decoder to reconstruct the original projected feature, as shown in Equation (4).

$$\hat{F}_m = D([S_m; C_m]) \tag{4}$$

where $D(\cdot)$ denotes the modality-specific decoder, and $[\cdot;\cdot]$ denotes the feature concatenation operation. Detailed formulations of the reconstruction and cycle consistency losses are provided in Appendix B. Optimizing this objective ensures that modality-invariant representations preserve unified sentiment semantics while maintaining strong discriminative power across categories, as shown in Equation (9).

$$\mathcal{L}_{FD} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{cycle} + \lambda_2 \mathcal{L}_{ort} + \lambda_3 \mathcal{L}_{sim} \tag{9}$$

In summary, the decorrelation constraint, reconstruction supervision, cycle consistency, and semantic alignment objective jointly govern the feature disentanglement process. This multi-faceted optimization enables our framework to generate structured representations that retain complete se-

mantic information while achieving clean decoupling. These optimized representations are subsequently fed into the Dual-Branch Masked Representation Enhancement (DBMRE) module for targeted refinement.

This reconstruction loss explicitly minimizes the discrepancy between the reconstructed and original features, mitigating semantic degradation caused by feature disentanglement.

Furthermore, to verify whether the reconstructed features successfully retain modality-specific characteristics, the modality-specific encoder is fed for re-encoding, as shown in Equation (6).

$$\hat{S}_m = E_s(\hat{Z}_m) \tag{6}$$

$$\mathcal{L}_{cycle} = \sum_m \|\hat{S}_m - S_m\|_2^2 \tag{7}$$

In tandem, we impose a cycle consistency constraint to guarantee representation stability, shown in Equation (7). This cycle consistency objective encourages the reconstructed features to preserve discriminative modality-specific traits without drifting from their original manifold.

In addition, modality-invariant representations extracted from different modalities should exhibit consistent semantic distributions. To achieve this, we introduce a margin-based semantic alignment objective. Specifically, representations sharing identical sentiment labels are pulled closer, whereas those from different sentiment categories are separated by a predefined margin, as shown in Equation (8).

$$\mathcal{L}_{sim} = \frac{1}{N}\sum_{i=1}^{N} \max(0, m + d(C_i, C_i^+) - d(C_i, C_i^-)) \tag{8}$$

where $d(\cdot)$ denotes the distance metric, and m denotes the margin parameter.

B.2 DBMRE Details

Guided by M_{ij}^{intra}, the masked representation learning is executed via a Transformer encoder, as shown in Equation (11).

$$S' = \text{Transformer}(S, M_{\text{intra}}) \tag{11}$$

The enhanced representations are subsequently partitioned back into their respective modality boundaries: $\tilde{S}_l, \tilde{S}_v, \tilde{S}_a$, and integrated with the original features via residual learning, as shown in Equation (12).

$$\tilde{S}_m = S_m + \lambda S'_m \tag{12}$$

where λ is a learnable scaling coefficient. By isolating information flow within each modality, this strategy effectively purifies modality-specific features, suppressing cross-modal noise and boosting robustness.

Subsequently, the Transformer encoder is employed to obtain the enhanced multimodal representations, as shown in Equation (15).

$$C' = \text{Transformer}(C, M_{\text{inter}}) \tag{15}$$

The output is partitioned back into respective modality branches: $\tilde{C}_l, \tilde{C}_v, \tilde{C}_a$, and preserved via residual learning, as shown in Equation (16).

$$\tilde{C}_m = C_m + \lambda C'_m \tag{16}$$

where λ is a learnable fusion parameter. This explicit elimination of intra-modal paths forces the model to concentrate entirely on cross-modal semantic alignment, yielding a more coherent shared semantic space.

B.3 ACIG Details

Taking the textual modality as an example, the attention outputs from the text-speech and text-visual interaction branches are concatenated to construct the cross-modal interaction features as shown in Equations (17) and (18).

$$H_{la} = \text{Trans}_{la}(\tilde{S}_l, \tilde{S}_a) \tag{17}$$

$$H_{lv} = \text{Trans}_{lv}(\tilde{S}_l, \tilde{S}_v) \tag{18}$$

where $[\cdot;\cdot]$ denotes the concatenation operation along the channel dimension. Because interaction features inherently contain temporal dynamics distributed across multiple time tokens, estimating channel importance directly from individual token representations can lead to unstable weight estimation.

Following this reweighting process, critical interaction responses are further enhanced, while noisy channels are effectively suppressed. Ultimately, the gated interaction features are fed into the multimodal memory Transformer to capture higher-order temporal dependencies, as formulated in Equation (22).

$$Z_l = \mathrm{MemTrans}(\tilde{H}_l) \tag{22}$$

For simplicity, the textual modality is used here as an illustrative example; the identical processing pipeline applies seamlessly to the visual and speech modalities.

B.4 Knowledge Distillation Details

Although the DBMRE and ACIG modules substantially advance multimodal representation learning from the standpoints of representation enhancement and interaction modeling, optimizing solely against the final prediction objective fails to provide sufficient supervision for intermediate representations. Specifically, modality-invariant representations and cross-modal interaction representations encode complementary semantic knowledge at distinct hierarchical levels. In the absence of explicit supervision, these intermediate features may progressively forfeit discriminative semantic details during end-to-end optimization, thereby constraining the model's overall representation capacity.

To address this limitation, we propose a multi-level knowledge distillation strategy that supplies auxiliary supervision across different representation spaces through homogeneous and heterogeneous knowledge distillation branches. Furthermore, we introduce a dictionary-guided contrastive learning objective to bolster semantic consistency among reconstructed multimodal representations. By jointly optimizing prediction-level supervision, representation-level knowledge distillation, and cross-modal semantic alignment objectives, our framework learns more discriminative and robust multimodal representations.

B.4.1 Homogeneous Knowledge Distillation

Modality-invariant representations are designed to capture shared semantic information across disparate modalities. To preserve the discriminative capacity of these shared representations, we introduce an auxiliary homogeneous knowledge distillation branch. Specifically, the modality-invariant features of each modality are first flattened and subsequently mapped into a compact latent space via a learnable projection head, as defined in Equation (23).

$$r_m = f_{\mathrm{proj}}(\mathrm{Flatten}(c_m)), \quad m \in \{l, v, a\} \tag{23}$$

where c_m denotes the enhanced modality-invariant representation of modality m from DBMRE, and $f_{\mathrm{proj}}(\cdot)$ represents a projection network composed of two fully connected layers augmented with a residual connection.

Subsequently, the projected representations are transformed into auxiliary sentiment predictions, as expressed in Equation (24).

$$\hat{y}_m^{\mathrm{homo}} = f_{\mathrm{cls}}(r_m) \tag{24}$$

where $f_{\mathrm{cls}}(\cdot)$ denotes the classification prediction head.

To preserve the semantic topology among modality-invariant representations, the homogeneous knowledge distillation objective accounts for three dimensions: prediction consistency, representation alignment, and structural regularization. The optimization objective is defined as Equation (25):

$$\mathcal{L}_{\mathrm{homo}} = \mathcal{L}_{\mathrm{logit}}^{\mathrm{homo}} + \mathcal{L}_{\mathrm{repr}}^{\mathrm{homo}} + \mathcal{L}_{\mathrm{reg}}^{\mathrm{homo}} \tag{25}$$

where $\mathcal{L}_{\mathrm{logit}}^{\mathrm{homo}}$ constrains the consistency between the auxiliary predictions and the ground-truth sentiment labels, as formulated in Equation (26).

$$\mathcal{L}_{\mathrm{logit}}^{\mathrm{homo}} = \sum_m |\hat{y}_m^{\mathrm{homo}} - y| \tag{26}$$

Furthermore, $\mathcal{L}_{\mathrm{repr}}^{\mathrm{homo}}$ preserves the semantic relationships among modality-invariant representations, while $\mathcal{L}_{\mathrm{reg}}^{\mathrm{homo}}$ maintains structural dependencies during graph-based knowledge transfer.

By incorporating this homogeneous knowledge distillation mechanism, modality-invariant representations from various modalities maintain coherent sentiment-relevant semantics prior to multimodal fusion, thereby enhancing the stability and discriminative power of the shared representations.

B.4.2 HETEROGENEOUS KNOWLEDGE DISTILLATION

Unlike its homogeneous counterpart, heterogeneous knowledge distillation focuses primarily on preserving complementary insights harvested through cross-modal interaction. Following adaptive interaction gating and multimodal memory modeling, each interaction representation successfully integrates sentiment cues from multiple modalities. Nevertheless, unguided interaction-oriented representations remain susceptible to a deficit in semantic discriminative power.

To mitigate this, we introduce a heterogeneous knowledge distillation branch. Initially, the interaction representations are mapped into a compact latent space, as shown in Equation (27).

$$r_m^{\text{hetero}} = g_{\text{proj}}(Z_m), \quad m \in \{l, v, a\} \tag{27}$$

where Z_m represents the output representation of the multimodal memory Transformer for modality m, and $g_{\text{proj}}(\cdot)$ denotes the heterogeneous projection head.

The projected interaction representations are subsequently transformed into auxiliary sentiment predictions, as shown in Equation (28).

$$\hat{y}_m^{\text{hetero}} = g_{\text{cls}}(r_m^{\text{hetero}}) \tag{28}$$

where $g_{\text{cls}}(\cdot)$ denotes the heterogeneous prediction head.

The heterogeneous knowledge distillation objective is defined as Equation (29).

$$\mathcal{L}_{\text{hetero}} = \mathcal{L}_{\text{logit}}^{\text{hetero}} + \mathcal{L}_{\text{repr}}^{\text{hetero}} + \mathcal{L}_{\text{reg}}^{\text{hetero}} \tag{29}$$

where the cross-entropy term supervises the consistency of auxiliary predictions derived from interaction-aware representations, $\mathcal{L}_{\text{logit}}^{\text{hetero}}$ is used to supervise the consistency between the auxiliary predictions generated from the interaction-aware representations, while $\mathcal{L}_{\text{repr}}^{\text{hetero}}$ is used to preserve the semantic structure encoded during cross-modal interaction.

B.5 DICTIONARY-GUIDED CONTRASTIVE LEARNING

Although homogeneous and heterogeneous knowledge distillation optimize representation learning via predictive and structural supervision, the reconstructed semantic representations still necessitate an explicit cross-modal alignment mechanism. To this end, we introduce a dictionary-guided contrastive learning objective to reinforce consistency across reconstructed multimodal representations.

Specifically, the reconstructed representations from all modalities are first aggregated: $R = [R_l, R_v, R_a]$, where R_m denotes the reconstructed semantic representation from Section 3.2. For each anchor representation, samples from other modalities sharing the identical sentiment label are treated as positive pairs, whereas those with differing sentiment labels constitute negative pairs. The contrastive learning objective is formalized in Equation (30).

$$\mathcal{L}_{\text{ctr}} = \frac{1}{N} \sum_i \max(0, \mu - \text{sim}(r_i, r_i^+) + \text{sim}(r_i, r_i^-)) \tag{30}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, r_i^+ and r_i^- denote the sets of positive and negative samples, respectively, and μ denotes the margin parameter.

By jointly considering homogeneous and heterogeneous reconstructed representations, the final dictionary-guided contrastive learning objective is defined as Equation (31).

$$\mathcal{L}_{DM} = \lambda_h \mathcal{L}_{\text{ctr}}^{\text{homo}} + \lambda_e \mathcal{L}_{\text{ctr}}^{\text{hetero}} + \lambda_c \mathcal{L}_{\text{cons}} \tag{31}$$

where the consistency constraint is introduced to maintain semantic stability between the original reconstructed representations and the enhanced reconstructed representations, as shown in Equation (32).

$$\mathcal{L}_{\text{cons}} = \frac{1}{3} \sum_{m \in \{l,v,a\}} \|R_m - R_m^{\text{enh}}\|_2^2 \tag{32}$$

Through dictionary-guided contrastive learning, the reconstructed representations from different modalities maintain sentiment-aware semantic consistency while preserving complementary sentiment information across modalities.

B.6 Fusion Prediction

Upon deriving the modality-invariant and interaction-aware representations under the multi-level supervision framework, the model integrates all complementary representations for final sentiment prediction.

Specifically, the interaction representations, shared semantic representations, and dictionary-reconstructed semantic prototypes are concatenated to construct the final multimodal representation: $\mathbf{h} = [\mathbf{h}_l^{\text{inter}}; \mathbf{h}_v^{\text{inter}}; \mathbf{h}_a^{\text{inter}}; \mathbf{c}^{\text{fusion}}; \mathbf{d}_l^{\text{homo}}; \mathbf{d}_v^{\text{homo}}; \mathbf{d}_a^{\text{homo}}; \mathbf{d}_l^{\text{hetero}}; \mathbf{d}_v^{\text{hetero}}; \mathbf{d}_a^{\text{hetero}}]$, where $\mathbf{c}^{\text{fusion}}$ denotes the fused modality-invariant representation, while $\mathbf{d}_m^{\text{homo}}$ and $\mathbf{d}_m^{\text{hetero}}$ denote the reconstructed semantic features obtained from the homogeneous and heterogeneous knowledge distillation branches, respectively. Subsequently, a multi-layer perceptron equipped with a residual connection refines the fused representation, as shown in Equation (33).

$$\hat{\mathbf{h}} = \mathbf{h} + \text{MLP}(\mathbf{h}) \tag{33}$$

Ultimately, the sentiment prediction results are generated via a linear prediction layer, as shown in Equation (34).

$$\hat{y} = W\hat{\mathbf{h}} + b \tag{34}$$

where W and b denote the learnable prediction parameters. The primary prediction objective is optimized by minimizing the sentiment regression loss, as shown in Equation (35).

$$\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{35}$$

where y_i and $\hat{y}_i$ denote the ground-truth sentiment labels and predicted sentiment scores, respectively.

Therefore, the overall knowledge distillation objective is defined as shown in Equation (36).

$$\mathcal{L}_{KD} = \mathcal{L}_{\text{homo}} + \mathcal{L}_{\text{hetero}} \tag{36}$$

Through the multi-level knowledge transfer mechanism, the proposed framework simultaneously preserves semantic consistency shared across modalities, complementary cross-modal interaction information, and graph structure-guided relational knowledge. Ultimately, the framework generates more discriminative multimodal sentiment representations, providing a more effective feature foundation for sentiment prediction.

C Datasets and Implementation Details

C.1 Datasets

To comprehensively evaluate the effectiveness of the proposed framework, we conduct extensive experiments on two widely adopted benchmark datasets for multimodal sentiment analysis: CMU-MOSI [Zadeh et al., 2016] and CMU-MOSEI [Zadeh et al., 2018b]. Both datasets comprise three heterogeneous modalities—text, speech, and visual information—and provide continuous sentiment intensity annotations, serving as standard testbeds for multimodal affective computing.

CMU-MOSI consists of 2,199 opinion video segments collected from YouTube, featuring 93 videos across 89 distinct speakers. Each segment is annotated with a continuous sentiment intensity score ranging from $[-3, +3]$. Following the standard experimental split, we utilize 1,284 samples for training, 229 for validation, and 686 for testing.

CMU-MOSEI is a significantly larger benchmark dataset, containing 3,228 videos and 22,856 available utterances from 1,000 speakers across 250 topics. Similar to CMU-MOSI, each utterance is labeled with a continuous sentiment intensity score within $[-3, +3]$. The detailed statistics of both datasets are summarized in Table 4. Following standard evaluation conventions, we focus primarily on continuous sentiment intensity prediction and binary/seven-class classification derived from these scores.

C.2 Evaluation Metrics

Following standard protocols in multimodal sentiment analysis, we evaluate model performance using both regression and classification metrics:

- **Regression Metrics:** Mean Absolute Error (MAE) (lower is better) and Pearson Correlation Coefficient (Corr) (higher is better).

Table 4: Dataset statistics

Dataset	Number of Videos	Number of Segments/Utterances	Sentiment Labels	Train	Val	Test	Speakers	Topics
CMU-MOSI [Zadeh et al., 2016]	93	2199	$[-3, +3]$	1284	229	686	89	—
CMU-MOSEI [Zadeh et al., 2018b]	3228	22856	$[-3, +3]$	16326	1871	4659	1000	250

Note: Train, Val, and Test denote the training, validation, and test sets, respectively. For CMU-MOSEI, although discrete emotion labels are also available, we strictly utilize continuous sentiment intensity labels to align with standard multimodal sentiment benchmarks.

- **Classification Metrics:** Binary Classification Accuracy (Acc-2), F1-Score (F1), and Seven-Class Classification Accuracy (Acc-7).

For binary classification, sentiment intensity scores > 0 are designated as positive and < 0 as negative. For seven-class classification, continuous scores are mapped to $\{-3, -2, -1, 0, 1, 2, 3\}$. All experimental results are reported as the mean over multiple independent runs.

C.3 Experimental Settings and Implementation Details

For the textual modality, we employ pre-trained BERT to extract contextual features, while visual and speech modalities utilize pre-extracted low-level descriptors, all projected into a shared space via three independent 1D convolutions. We adopt a word-level aligned sequence length of 50 across all modalities for both datasets. The model is optimized end-to-end using the Adam optimizer with a ReduceLROnPlateau scheduler for up to 50 epochs. A linear warm-up strategy increases the dictionary-guided contrastive loss weight from 0 to 0.05 over the first 10 epochs. The multi-objective loss combines L1 task regression, auxiliary distillation losses (each weighted at 0.05), and hierarchical disentanglement constraints, supplemented by an Intra/Inter branch L2 regularization of 1×10^{-4} and a contrastive margin of 0.05. The relevant parameters are summarized in Table 5.

Table 5: Training and optimization parameters of the model

Category	**Parameter**	**Setting**
Dataset	Dataset	CMU-MOSI, CMU-MOSEI
Input Modalities	Modalities	Text, Visual, Audio
Text Encoder	Text Encoder	Pre-trained BERT
Feature Projection	Projection	3 independent 1D convolutions
Alignment Setting	Aligned Sequence Length	Text = 50, Visual = 50, Audio = 50
Optimization	Optimizer	Adam
Learning Rate Scheduling	Scheduler	ReduceLROnPlateau
Maximum Training Epochs	Max Epochs	50
Dictionary Contrastive Loss	dm_weight	Epoch 1–10: $0 \rightarrow 0.05$, linear warm-up; 0.05 thereafter
Task Loss	Task Loss	L1: Main + Homo. Aux. + Hetero. Aux. + Shared Repr. Prediction
Homogeneous Graph Distillation	$\mathcal{L}_{\text{homo}}$	0.05
Heterogeneous Graph Distillation	$\mathcal{L}_{\text{hetero}}$	0.05
Disentanglement-related Loss	Recon. / Cycle / Sim. / Orth.	$(\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{cycle}} + (\mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{ort}}) \times 0.1) \times 0.1$
Module Regularization	Intra/Inter L2	1×10^{-4}
Contrastive Learning	Margin	0.05

Note: dm_weight denotes the weight of the dictionary-guided contrastive loss. Intra/Inter L2 denotes the L2 regularization for the intra-/inter-modal representation branches.

D Additional Experimental Results

D.1 Detailed Comparison with Existing Methods

To comprehensively evaluate the performance of the proposed framework, we compare it with several representative methods on the CMU-MOSI and CMU-MOSEI datasets. The compared methods can be broadly divided into three categories: (1) classical multimodal fusion methods, including TFN [Zadeh et al., 2017], LMF [Liu et al., 2018], MFN [Zadeh et al., 2018a], and MulT [Tsai et al., 2019]; (2) feature disentanglement and alignment methods, including MISA [Hazarika et al., 2020] and Self-MM [Yu et al., 2021]; and (3) recent attention- and gating-enhanced methods, including MTAG [Yang et al., 2021].

To ensure the traceability of the comparison results, except for DHMD and our method, the results of all other baselines reported in Table 1 are directly adopted from their respective original publications. To further evaluate the effectiveness of the proposed modules on a strong recent baseline, we select the latest Decoupled Hierarchical Multimodal Distillation (DHMD) [Li et al., 2026] as the primary comparison model. DHMD is highly relevant to our research problem, as it disentangles multimodal features into modality-invariant and modality-specific representations and combines graph distillation with cross-modal dictionary matching to achieve hierarchical knowledge transfer.

Following the original DHMD paper and its publicly available implementation, we strictly reproduce its core architecture, training objectives, and experimental settings. The reproduced results are denoted as "DHMD". On this basis, we only add the proposed Dual-Branch Masked Representation Enhancement (DBMRE) module and Adaptive Cross-modal Interaction Gating (ACIG) module, while keeping the feature extraction, disentanglement, knowledge distillation, and dictionary matching mechanisms unchanged, thereby constructing the final model, denoted as "Ours". This incremental design aims to rigorously isolate the performance gains contributed by DBMRE and ACIG and avoid potential confounding factors caused by changes in the overall model architecture. The comparison results are presented in Table 1.

Several observations can be drawn from Table 1. First, compared with classical early fusion methods (TFN, LMF, and MulT), the proposed framework achieves substantial performance improvements on CMU-MOSI, with gains of 4.26–6.96 percentage points in Acc-2 and 12.10–15.20 percentage points in Acc-7, while reducing MAE by 0.214–0.261. These results indicate that simple feature concatenation or fixed interaction strategies are insufficient to capture fine-grained cross-modal semantics. In contrast, the progressive modeling process of "disentanglement to dual-branch enhancement to adaptive interaction selection" enables the proposed framework to more effectively exploit complementary sentiment cues from heterogeneous modalities.

Second, compared with representative disentanglement-based methods (MISA and Self-MM), the proposed method further improves Acc-2 by 1.76–2.66 percentage points and reduces MAE to 0.711. This improvement can be mainly attributed to two factors. First, DBMRE performs targeted enhancement of disentangled representations through intra-modal and inter-modal masking, addressing the limitation of existing disentanglement methods that emphasize separation while paying insufficient attention to subsequent representation optimization. Second, ACIG dynamically selects interaction channels according to their semantic context, thereby avoiding the amplification of noise caused by equal-weight aggregation. The synergistic effect of these two components further validates the two core design motivations introduced in Section 1.

Finally, compared with recent attention- and gating-based methods (MTAG and GraphCAGE), the proposed framework achieves competitive performance across multiple evaluation metrics, particularly benefiting from representation-aware masking and adaptive interaction gating, indicating that the proposed representation-aware masking and adaptive gating mechanisms are more effective than simple attention enhancement or graph-based propagation strategies. Overall, the proposed framework achieves competitive performance on mainstream benchmarks, demonstrating the effectiveness of the "representation enhancement + interaction selection" paradigm for multimodal sentiment analysis.

D.2 Detailed Ablation Analysis

Effectiveness of Dual-Branch Masking (DBMRE). As shown in Table 2, introducing only the DBMRE module to the baseline boosts Acc-2 and F1 to 85.06%, marking a substantial improvement of 1.37 percentage points alongside a reduced MAE. This confirms that targeted intra-modal masking for modality-specific features and inter-modal masking for modality-invariant representations significantly elevate the quality and discriminative power of disentangled features.

Synergistic Impact of Adaptive Gating (ACIG). Interestingly, applying ACIG alone results in marginal regression degradation (e.g., MAE increases to 0.737). This intuitive phenomenon reveals a crucial insight: performing interaction selection without prior feature enhancement may amplify noise or starve downstream classifiers of reliable semantic cues. However, when integrated alongside DBMRE, the full framework achieves peak performance, securing an outstanding Acc-7 of 47.90% and the lowest MAE of 0.711. This underscores the precise complementarity of our design: DBMRE first purifies and strengthens the representation foundation, enabling ACIG to accurately distill and amplify informative cross-modal interaction paths.

Superiority over Static Linear Gating. Replacing the context-dependent ACIG with static linear gating yields performance degradation across all metrics (Table 3, row 1), dropping Acc-2 to 83.46% and elevating MAE to 0.729. This validates our core hypothesis: gating coefficients must be dynamically generated based on the runtime semantic context of each interaction branch to effectively handle sample-wise heterogeneity.

Comparison with Alternative Masking and Interaction Strategies. Replacing DBMRE with random feature masking or substituting cross-modal interaction with GCN-based message passing leads to sharp drops in performance (Table 3). This verifies that our structured attention constraints and representation-aware masking provide much more targeted preservation of affective semantics than heuristic or random alternatives.

Analysis of Single-Branch Variants. Retaining solely the modality-specific (S) branch yields a strong Acc-7 of 47.38%, whereas retaining only the modality-invariant (C) branch causes a marked performance drop. This indicates that modality-specific cues play a particularly decisive role in fine-grained sentiment intensity prediction. Nonetheless, the full dual-branch design achieves the optimal balance across all evaluation metrics by harmonizing invariant and specific perspectives.

E Training Dynamics

E.1 Further Analysis

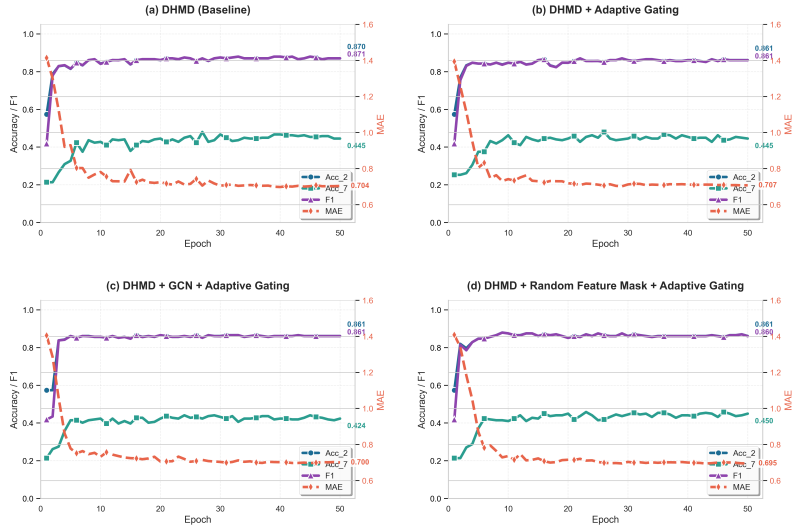

Figure 4: Performance dynamics of different interaction/masking variants during training on the CMU-MOSI validation set: (a) DHMD (Reproduced); (b) baseline + adaptive gating (ACIG); (c) baseline + GCN-based cross-modal message passing + ACIG; and (d) baseline + random feature masking + ACIG. Acc-2, Acc-7, and F1 (left axis) and MAE (right axis) are plotted against the training epochs in each subfigure.

The modality-specific representations (S branch) and modality-invariant representations (C branch) produced by the feature disentanglement module serve different semantic functions, and DBMRE therefore employs distinct masking enhancement strategies for the two branches. To unpack their respective contributions, Figure. 6 compares the training curves of the S and C branches when only one branch is retained, with both variants combined with ACIG, across the four evaluation metrics.

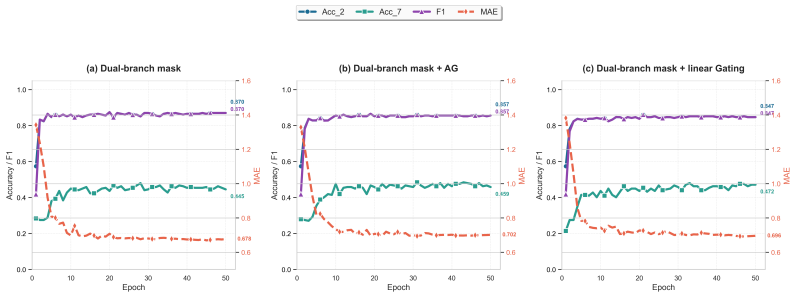

Figure 5: Training dynamics of the combinations of dual-branch masking and different gating strategies: (a) dual-branch masking only (DBMRE); (b) DBMRE + adaptive gating (full model); and (c) DBMRE + linear gating.

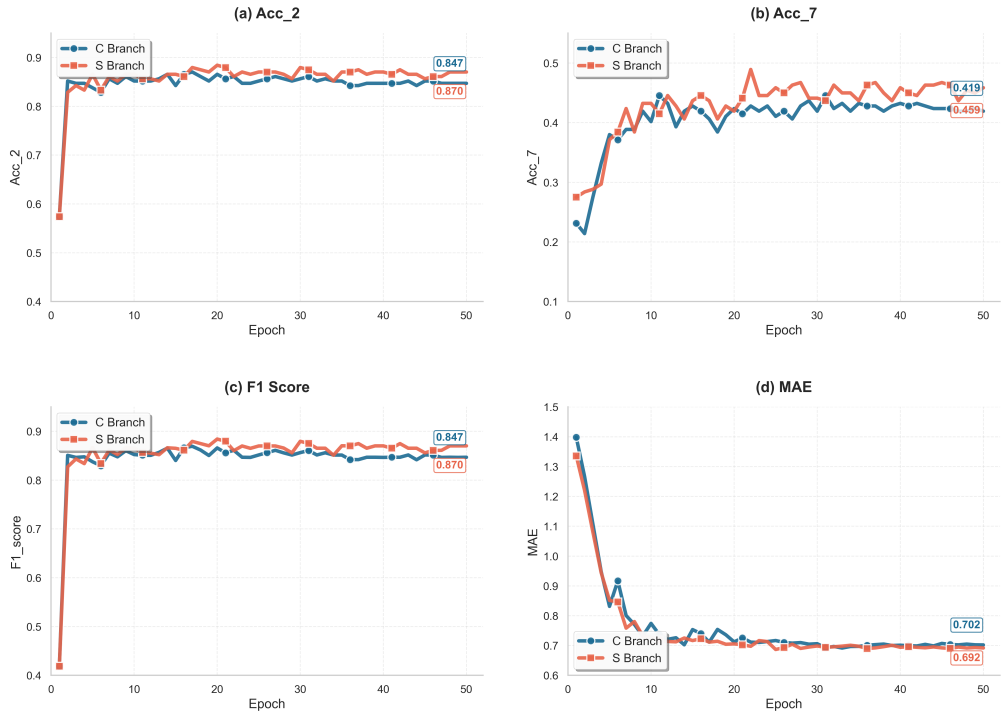

Figure 6: Training curves of the modality-specific branch (S) and modality-invariant branch (C) across four evaluation metrics on the CMU-MOSI validation set: (a) Acc-2; (b) Acc-7; (c) F1; and (d) MAE.

As illustrated in Figure. 6, several critical insights emerge:

Fine-grained sentiment relies predominantly on modality-specific cues. The training trajectories of the S branch consistently outperform those of the C branch across all metrics. Specifically, the S branch achieves higher final Acc-2 $(84.7\% - 87.0\%)$ and Acc-7 $(41.9\% - 45.9\%)$, alongside a lower MAE $(0.692 - 0.702)$. This confirms that modality-specific attributes carry decisive discriminative signals for fine-grained sentiment intensity regression, corroborating our design rationale in Section 3.1 that disentangled representations require targeted, branch-specific purification.

The optimization process exhibits convergence behavior and optimization difficulty. The S branch establishes a performance advantage early in training (within the first 10 epochs) and maintains it throughout. In contrast, the C branch exhibits pronounced fluctuations. This divergence stems from their semantic nature: modality-specific features are directly sharpened via intra-modal masked attention, whereas invariant features rely on cross-modal semantic consistency, making cross-modal alignment inherently more challenging and slower to stabilize.

It exhibits complementary synergy. Despite the weaker standalone performance of the C branch, the full dual-branch framework (combining both S and C) surpasses the S-only variant on overall metrics (as shown in Table 2). This confirms that shared semantics supplied by the C branch act as a vital regularizer and semantic complement to modality-specific cues, validating the necessity of the dual-branch DBMRE architecture.

F Computational Complexity Analysis

F.1 Computational Complexity Analysis

To evaluate the computational efficiency of the proposed framework, we analyze the theoretical time complexity of DBMRE, ACIG, the cross-modal Transformer, and the multimodal memory Transformer, comparing them with representative Transformer-based baselines. Let the sequence length of each modality be T, the unified embedding dimension be d, and the number of attention heads be H. Under the aligned setting adopted in this work, the sequence lengths of the three modalities are uniformly set to $T = 50$.

(1) DBMRE incorporates intra- and inter-modal masked self-attention over concatenated sequences. Since the concatenated sequence length is $3T$, the primary computation arises from attention matrix construction, yielding a complexity of $\mathcal{O}((3T)^2 d) = \mathcal{O}(9T^2 d)$. Because masking merely reconfigures attention connections without introducing sequence-length-dependent parameters, DBMRE preserves the asymptotic quadratic complexity order $\mathcal{O}(T^2 d)$.

(2) ACIG first applies global average pooling over the sequence dimension with a complexity of $\mathcal{O}(Td)$, followed by two lightweight fully connected layers for channel-wise gating projection with a complexity of $\mathcal{O}(d^2)$. For all three modalities, the total complexity is bounded by $\mathcal{O}(3(Td + d^2))$. Crucially, ACIG bypasses dense all-pair sequence attention, introducing only a low-order additional overhead.

(3) Cross-modal Transformer and Memory Transformer. During the cross-modal interaction stage, each interaction branch employs a Transformer to model dependencies among different modalities, with the main computational complexity of the attention operation given by $\mathcal{O}(T^2 d)$. Subsequently, the multimodal memory Transformer further models the gated and reweighted interaction representations, with its main computational complexity likewise given by $\mathcal{O}(T^2 d)$.

(4) Overall Computational Complexity. The dominant term of the proposed framework remains governed by multi-head self-attention at $\mathcal{O}(T^2 d)$, while ACIG and DBMRE introduce only lower-order terms. Thus, the proposed framework maintains identical asymptotic complexity $\mathcal{O}(T^2 d)$ to mainstream Transformer-based baselines while significantly boosting representation quality and interaction selectivity.

G Discussion and Limitations

G.1 Discussion and Limitations

Empirical Alignment with Methodological Motivations. The experimental findings establish a direct, watertight correspondence with the core motivations outlined in Section 1 and Section 3. The motivation that disentangled representations suffer from information loss and require explicit post-hoc enhancement is validated by the 1.37% Acc-2 gain achieved by DBMRE alone (Table 2). The motivation that uniform feature propagation across cross-modal interaction branches introduces noise is substantiated by the performance drop of linear gating and the superior stability of ACIG (Table 3 and Figure. 5). The full model peak performance across Acc-7 and MAE proves the efficacy of our progressive paradigm.

Limitations. Despite its competitive empirical performance, our study presents several limitations that invite future exploration. The framework relies heavily on pre-trained BERT features for the textual stream. Its adaptability and robustness in low-resource regimes or settings devoid of large-scale pre-trained models require further investigation. Although asymptotic complexity remains quadratic, the introduction of dual-branch masking and gating modules marginally increases architectural complexity and hyperparameter sensitivity (e.g., masking ratios and projection dimensions), posing challenges for extreme real-time edge deployment. Current evaluations focus primarily on English benchmarks (CMU-MOSI and CMU-MOSEI). Generalizability to multilingual corpora (e.g., Chinese affective benchmarks such as CH-SIMS) remains to be tested. While ACIG

mitigates feature noise, the model's resilience under severe real-world challenges—such as missing modalities or extreme adversarial noise—warrants deeper study.

Future Work. Based on the above analysis, future research will focus on several directions. First, we will investigate more lightweight and efficient masking and gating mechanisms and explore knowledge distillation or structured pruning to reduce computational complexity and facilitate practical deployment. Second, we will improve the model's adaptive capability under missing-modality and severe-noise conditions and incorporate visualization and causal analysis techniques to enhance the interpretability of the decision-making process. Third, we will extend the proposed disentanglement enhancement and adaptive interaction mechanisms to multilingual, cross-domain, and conversational sentiment understanding tasks, while exploring their integration with multimodal large language models to further improve understanding and reasoning in complex sentiment scenarios.

G.2 Conclusion and Future Work

In this work, we proposed a novel disentangled representation learning framework for multimodal sentiment analysis, centered on the Dual-Branch Masked Representation Enhancement (DBMRE) and Adaptive Cross-modal Interaction Gating (ACIG) modules. Addressing the critical limitations in existing literature—namely, the neglect of post-separation representation optimization and the susceptibility of cross-modal transformers to noisy dependencies—DBMRE effectively boosts intra-modal discriminative semantics and inter-modal consistency through dedicated masking branches. Building upon these enhanced representations, ACIG dynamically refines cross-modal interactions via global context-aware channel-wise weighting. Furthermore, the integration of multi-level knowledge distillation and dictionary-guided contrastive learning provides robust auxiliary supervision, ensuring high semantic coherence. Extensive experiments across benchmark datasets, supported by rigorous ablation studies, validate the synergistic contributions of DBMRE and ACIG, demonstrating state-of-the-art performance in both classification and regression metrics.

Despite these advances, several limitations remain. First, the framework's reliance on frozen or fine-tuned pre-trained text encoders establishes an upper bound on feature extraction quality, which propagates through the subsequent disentanglement and interaction pipelines. Second, the incorporation of dual-branch masking and adaptive gating introduces non-trivial computational overhead, challenging deployment in resource-constrained or real-time edge environments. Finally, while evaluations on standard benchmarks are comprehensive, the model's robustness under real-world distribution shifts, severe modality conflicts, and noisy data streams warrants deeper investigation.

To address these challenges, future work will pursue three primary directions: We will explore parameter-efficient masking strategies, structured pruning, and lightweight distillation techniques to drastically reduce computational complexity for practical deployment. We aim to enhance the framework's resilience against missing modalities and high-noise conditions, while incorporating visualization and causal inference tools to decode the decision-making rationale of the disentanglement and gating mechanics. We will extend the DBMRE-ACIG paradigm to multilingual, cross-domain, and multi-turn conversational sentiment understanding, while investigating its seamless integration with Multimodal Large Language Models (MLLMs) to unlock advanced reasoning in complex affective computing scenarios.