SLaD: Sub-modal Label-aware Disentanglement for Multimodal Sentiment Analysis

Anonymous authors
Paper under double-blind review

Abstract

Multimodal sentiment analysis (MSA) requires integrating heterogeneous information effectively while addressing inconsistent emotional cues across modalities. However, existing approaches often fail to disentangle modality-invariant and modality-specific representations, leading to suboptimal feature alignment and semantic entanglement, especially when emotional expressions differ across sub-modalities. To address this issue, we propose a Sub-modal Label-aware Disentanglement (SLaD) framework that enhances cross-modal representation learning through a sub-modal label similarity weighting mechanism. Specifically, SLaD defines three structural relationships among sub-modal labels and introduces a hybrid similarity function that integrates structural consistency with numerical similarity. This approach mitigates label noise and conflicts from heterogeneous modality information. We further introduce three complementary losses for joint optimization: (1) a modality contrastive loss that aligns modality-invariant features, (2) a modality repulsive loss that enhances the discriminability of modality-specific features, and (3) a multi-label contrastive loss that captures sub-modal emotional label correlations. Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS demonstrate that SLaD achieves state-of-the-art performance on both classification and regression tasks, demonstrating the effectiveness of sub-modal label-aware supervision and disentanglement for advancing multimodal sentiment understanding.

1 Introduction

The rise of social media and multimodal content (e.g., text, images, videos) Li et al. (2022) has made sentiment analysis more challenging and essential for tasks like retrieval, opinion tracking, diagnosis, and user feedback. Traditional single-modal sentiment analysis, which relies solely on text or audio, often fails to fully capture human emotions, as emotional expression in real-world scenarios is inherently multimodal Shoumy et al. (2020). In contrast, MSA integrates heterogeneous modalities to provide a more comprehensive understanding of human emotions.

Recent research has shown that leveraging complementary information across text, visual, and audio modalities significantly improves sentiment understanding Zubatyuk et al. (2019). However, the inherent heterogeneity of multimodal data still poses significant challenges, such as inconsistencies in sentiment expression across modalities Xiao et al. (2023), irrelevant information in heterogeneous modalities Li et al. (2025b), and limited availability of distinctive emotional cues in facial expressions Wu et al. (2024). Therefore, MSA models must capture inter-modal consistency while maintaining intra-modal coherence. To address these challenges, researchers have proposed methods that extract and integrate diverse emotional cues from multiple modalities, which generally fall into two paradigms: representation-based and interaction-based approaches Li et al. (2025a). Representation-based methods encode each modality using modality-specific subnetworks to generate global representations, often incorporating additional constraints to enhance representation quality Wang et al. (2024); Li et al. (2025b). In contrast, interaction-based methods design sophisticated architectures to capture fine-grained token-level interactions across modalities Li et al. (2025a); Jin et al. (2024). Most of these methods fail to account for multiple emotional labels and modality-specific distinctions, which hampers their ability to model inter-modal relationships under label inconsistency accurately. Figure 1 shows an example of label discrepancy, where positive sentiment in text co-occurs with negative sentiment in audio.

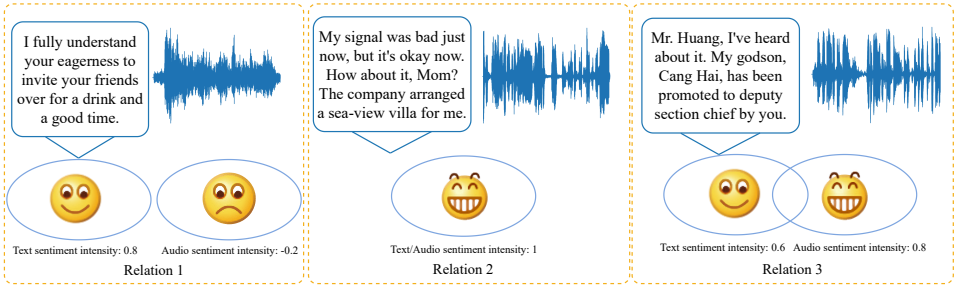


Figure 1: Relations between text and audio sub-modal labels: Relation 1 denotes non-overlapping labels, Relation 2 denotes fully co-occurring labels, and Relation 3 denotes partially co-occurring labels.

Supervised learning struggles with multi-label inconsistencies despite excelling at single-label tasks. This necessitates modeling implicit inter-modality relationships, especially across varying label co-occurrence patterns. Additional complexities arise from difficult-to-align cross-modal features.

To address these challenges, we propose a Sub-modal Label-aware Disentanglement (SLaD) framework that explicitly models both structural relationships, such as cases where one sub-modal label is ”0” or modal labels have opposite signs—and numerical relationships, defined by the similarity between sub-modal emotional labels, to guide cross-modal representation learning. SLaD defines three types of label-set relationships: non-overlapping, fully co-occurring, and partially co-occurring. To leverage these relationships, it introduces a hybrid similarity weighting mechanism that integrates structural consistency with numerical similarity. This mechanism ensures that only structurally comparable label pairs are aligned, effectively reducing semantic conflicts and mitigating noise introduced by inconsistent labels. To disentangle modality-invariant and modality-specific representations, SLaD further incorporates three complementary loss functions: (1) a modality contrastive loss to align shared emotional semantics, (2) a modality repulsive loss to enhance modality-specific discriminability and suppress cross-modal interference, and (3) a multi-label contrastive loss to exploit inter-label correlations for improved emotional association modeling. By jointly optimizing these components, SLaD effectively mitigates sentiment label inconsistency, enhances cross-modal affective alignment, and improves the overall performance of MSA. The main contributions can be summarized as follows:

- We propose a novel SLaD framework that leverages structural and numerical relationships among sub-modal labels to guide cross-modal alignment, effectively addressing label inconsistency and semantic conflicts.
- We introduce a modality contrastive loss and a modality repulsive loss to simultaneously align shared semantics and preserve modality-specific independence, enabling more robust representation learning.
- We design a multi-label contrastive loss, which dynamically adjusts similarity constraints based on label-set overlap, allowing the model to capture inter-label correlations and improve emotional association modeling.
- SLaD has conducted extensive experiments on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, demonstrating that the method outperforms baselines in both classification and regression tasks, validating its effectiveness in MSA.

2 Related Work

Multimodal sentiment analysis (MSA) exploits complementary cues from language, audio, and vision. Prior works have emphasized cross-modal fusion: Huang et al. (2024a) introduced a binding mechanism, Gan et al. (2024) mapped heterogeneous features into a shared space, and Xie et al. (2025) pursued deep fusion through dense layers. While effective, these methods often overlook modality conflicts and redundancy.

Contrastive learning has been widely applied for semantic alignment. Zong et al. (2023) maximized mutual information with an InfoNCE loss, Yang et al. (2023) enhanced intra-modal discrimination, and Fan et al. (2025) reduced cross-modal heterogeneity via multi-level objectives. However, most focus on global consistency and ignore the separation of modality-invariant and modality-specific signals, which limits robustness to ambiguous samples.

Multi-label learning further addresses label correlations. Li et al. (2024) built a multi-label detection module, Chen et al. (2024) added a label association loss, and Deng et al. (2023) balanced positive and negative predictions with a focus loss. Yet these approaches mainly capture global label dependencies while neglecting fine-grained sub-modal label inconsistencies.

In summary, existing studies have advanced fusion, alignment, and label modeling, but few integrate these perspectives. The lack of frameworks that disentangle invariant/specific features while accounting for sub-modal label discrepancies motivates our proposed approach.

3 METHODOLOGY

As an overview of the model, Figure 2 illustrates the architecture of SLaD.

3.1 PROBLEM DEFINITION

This study focuses on a multimodal sentiment analysis (MSA) task involving two modalities: text (t) and audio (a). We denote the modality set as $\mathcal{M} = \{t, a\}$, where $m \in \mathcal{M}$ represents a specific modality. The sentiment prediction $\hat{y}$ is a continuous value in either $[-1, 1]$ or $[-3, 3]$, with $\hat{y} > 0$, $\hat{y} = 0$, and $\hat{y} < 0$ indicating positive, neutral, and negative sentiment, respectively.

3.2 MULTIMODAL FEATURE EXTRACTION

To ensure optimal performance of SLaD across different modalities and languages, we employ language-specific pre-trained models tailored to each dataset's linguistic characteristics. For English datasets, we utilize RoBERTa Liu et al. (2019) and Data2Vec Baevski et al. (2022) to extract textual and acoustic features, respectively. For Chinese datasets, we adopt Chinese-RoBERTa Cui et al. (2020) and Chinese-HuBERT models to handle the language-specific nuances effectively. The extracted textual and acoustic modal features are represented as $X^t \in \mathbb{R}^{B \times l_t \times d}$ and $X^a \in \mathbb{R}^{B \times l_a \times d}$, respectively, where B denotes the batch size, l_t and l_a represent the sequence lengths for textual and acoustic modalities, and d indicates the feature dimensionality.

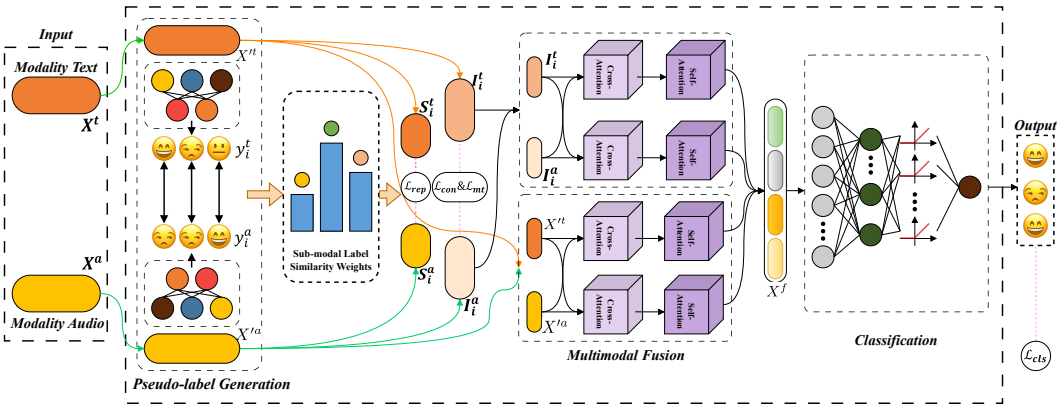


Figure 2: The overall architecture of SLaD consists of multimodal feature extraction, a modality-invariant sentiment representation, and multimodal fusion & classification.

3.3 PSEUDO-LABEL GENERATION

The two unimodal pseudo-label generation tasks, which share representations with the multimodal task, are addressed by first projecting these representations into a dedicated feature space. This projection mitigates dimensional discrepancies across modalities and ensures more consistent feature distributions. Linear regression is then applied to map the projected unimodal features to their corresponding pseudo-labels, producing predicted pseudo-labels for each unimodal task. These predictions serve as inputs to the Sub-modality Label-aware Similarity Modeling module, forming a robust foundation for cross-sub-modality similarity learning.

$$y_i^t = W_{l1}^{tT} X'^t + b_{l1}^t \tag{1}$$

$$y_i^a = W_{l1}^{aT} X'^a + b_{l1}^a \tag{2}$$

3.4 MODALITY-INVARIANT SENTIMENT REPRESENTATION

To facilitate effective modality-invariant sentiment representation, We first align the text features and audio features extracted by the pre-trained model in terms of sequence length, yielding unified representations $X'^t \in \mathbb{R}^{B\times l\times d}$ and $X'^a \in \mathbb{R}^{B\times l\times d}$. Subsequently, these aligned features undergo a disentanglement process that decomposes them into two complementary components: modality-invariant features $I_i^t \in \mathbb{R}^{B\times l\times d}$ and $I_i^a \in \mathbb{R}^{B\times l\times d}$, which capture shared semantic information across modalities, and modality-specific features $S_i^t \in \mathbb{R}^{B\times l\times d}$ and $S_i^a \in \mathbb{R}^{B\times l\times d}$, which preserve unique modal characteristics.

3.5 MULTIMODAL FUSION & CLASSIFICATION

Building upon the extracted modality-invariant sentiment features and inspired by Wu et al. (2024), we design a fusion network based on a cross-attention encoder. In this architecture, queries from one modality attend to keys and values from another, enabling the model to capture fine-grained inter-modal dependencies. Such cross-modal interactions facilitate a more effective integration of heterogeneous modalities, leading to a richer and more comprehensive representation, as illustrated below:

$$\mathrm{Attention}(\mathrm{Q}_{I_i^t}, \mathrm{K}_{I_i^a}, \mathrm{V}_{I_i^a}) = \mathrm{softmax}\left(\frac{\mathrm{Q}_{I_i^t}\mathrm{K}_{I_i^a}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}_{I_i^a} \tag{3}$$

$$\mathrm{Attention}(\mathrm{Q}_{I_i^a}, \mathrm{K}_{I_i^t}, \mathrm{V}_{I_i^t}) = \mathrm{softmax}\left(\frac{\mathrm{Q}_{I_i^a}\mathrm{K}_{I_i^t}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}_{I_i^t} \tag{4}$$

$$\mathrm{Attention}(\mathrm{Q}_{X'^t}, \mathrm{K}_{X'^a}, \mathrm{V}_{X'^a}) = \mathrm{softmax}\left(\frac{\mathrm{Q}_{X'^t}\mathrm{K}_{X'^a}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}_{X'^a} \tag{5}$$

$$\mathrm{Attention}(\mathrm{Q}_{X'^a}, \mathrm{K}_{X'^t}, \mathrm{V}_{X'^t}) = \mathrm{softmax}\left(\frac{\mathrm{Q}_{X'^a}\mathrm{K}_{X'^t}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}_{X'^t} \tag{6}$$

Specifically, the query, key, and value representations $\mathrm{Q}_{I_i^t}$, $\mathrm{K}_{I_i^t}$, and $\mathrm{V}_{I_i^t}$ are generated from the modality-invariant text features I_i^t, whereas $\mathrm{Q}_{I_i^a}$, $\mathrm{K}_{I_i^a}$, and $\mathrm{V}_{I_i^a}$ are generated from the modality-invariant audio features I_i^a. In parallel, $\mathrm{Q}_{X'^t}$, $\mathrm{K}_{X'^t}$, and $\mathrm{V}_{X'^t}$ are obtained from the text input X'^t, and $\mathrm{Q}_{X'^a}$, $\mathrm{K}_{X'^a}$, and $\mathrm{V}_{X'^a}$ are obtained from the audio input X'^a. To preserve temporal dynamics, the network further employs self-attention encoders that model temporal dependencies within the cross-modal features. This dual-attention architecture jointly captures cross-modal complementarity and intra-modal temporal coherence.

Finally, a position-wise feed-forward network, consisting of fully connected layers with ReLU activation, is applied to refine the feature representations at each time step.

3.6 SUB-MODALITY LABEL-AWARE SIMILARITY MODELING

To address the fact that labels across different modalities often exhibit inconsistent polarities or misaligned expression intensities, we propose a structure-aware similarity modeling mechanism for

sub-modal label perception. This mechanism comprehensively considers both set-theoretic relationships and numerical differences among sub-modal labels. As shown in Figure 3, we represent each relationship as $\mathcal{R}$. For example, $\mathcal{R}_1$ represents Relationship 1, where y_i^t and y_i^a denote the text sub-modal sentiment intensity label and the audio sub-modal sentiment intensity label, respectively. The three relationships are defined as follows:

$$\mathcal{R}_1 : \text{sign}(y_i^t) \neq \text{sign}(y_i^a) \vee y_i^t = 0 \vee y_i^a = 0 \tag{7}$$

$$\mathcal{R}_2 : y_i^t = y_i^a \tag{8}$$

$$\mathcal{R}_3 : \text{sign}(y_i^t) = \text{sign}(y_i^a) \wedge y_i^t \neq y_i^a \wedge y_i^t \neq 0 \wedge y_i^a \neq 0 \tag{9}$$

where $\mathcal{R}_1$ represents semantic incompatibility (opposite polarities or zero values), $\mathcal{R}_2$ denotes perfect label alignment, and $\mathcal{R}_3$ indicates partial consistency with intensity variations. Based on the above-defined relationships, this partitioning of structural relationships provides a systematic foundation for consistency modeling of multimodal labels, establishing the groundwork for subsequent similarity calculations.

Building upon the structural relationships, we further propose a hybrid label similarity function $\omega_i \in [0, 1]$ that fuses structural consistency and numerical similarity. This function is designed to calculate the label consistency weight for each sample at the sub-modal level. Through this mechanism, label alignment is ensured only when the semantic structures are consistent. The structural weight component is defined as:

$$\omega_i^{\text{structure}} = \begin{cases} 0 & \text{if } \mathcal{R}_1 \text{ holds} \\ 1 & \text{if } \mathcal{R}_2 \text{ holds} \\ 0.1 + 0.8 \cdot \frac{\min(|y_i^t|, |y_i^a|)}{\max(|y_i^t|, |y_i^a|) + \epsilon} & \text{if } \mathcal{R}_3 \text{ holds} \end{cases} \tag{10}$$

where $\epsilon = 10^{-8}$ prevents division by zero. For relationship $\mathcal{R}_3$, we introduce an adaptive weighting scheme based on relative intensity similarity, where labels with closer magnitudes receive higher structural weights within the range $[0.1, 0.9]$. This structural consistency modeling leverages a relationship-aware mechanism to ensure that only sub-modal sample pairs with consistent label semantic structures participate in the similarity calculation and supervision.

To maintain structural consistency, we also consider the numerical differences in sentiment labels. The specific rationale is as follows: if we simply determine that labels with identical polarity represent complete semantic equivalence, the intensity differences in sentiment expression will be ignored. This approach makes the supervision signal excessively coarse-grained and may forcibly align features that are not semantically similar. Given this consideration, we adopt a Gaussian kernel function to construct a similarity measurement mechanism sensitive to intensity differences, thereby accurately capturing the numerical variations of sentiment labels. The numerical similarity modeling is formulated as:

$$\omega_i^{\text{value}} = \exp\left(-\frac{(y_i^t - y_i^a)^2}{2\alpha^2}\right) \tag{11}$$

where α is the bandwidth parameter of the Gaussian kernel function, which determines the influence range of the kernel function. The Gaussian kernel function provides smooth similarity measurement, assigning higher weights to labels with smaller numerical discrepancies while implementing progressive supervision that avoids excessive penalties for minor expression inconsistencies. Based on

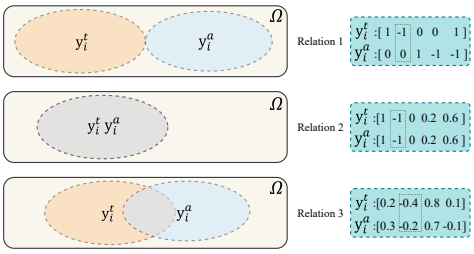


Figure 3: The three set relationships among sub-modal labels are exemplified by the text and audio sub-modal label in CH-SIMS. Ω represents the universal set containing all sub-modal label entities.

the definitions of structural consistency and numerical similarity mentioned above, the sub-modality label similarity weight is $\omega_i = \omega_i^{\text{structure}} \cdot \omega_i^{\text{value}}$, with final clamping to ensure $\omega_i \in [0, 1]$. For a batch size of B, the vector expression of the similarity weight is as follows:

$$\boldsymbol{\omega} = [\omega_1, \omega_2, \cdots, \omega_B]^{\mathrm{T}} \in [0, 1]^B \tag{12}$$

3.7 Modality Alignment and Disentanglement Modeling

To achieve semantic consistency modeling and modal heterogeneity enhancement in MSA, this study introduces two synergistic loss mechanisms in the representation learning of modality-invariant and modality-specific features: the modality contrastive loss $\mathcal{L}_{con}$ and the modality repulsive loss $\mathcal{L}_{rep}$. The former is used to align modality-invariant features from different modalities, making them express similar semantics; the latter encourages modality-specific features to maintain differences in the semantic space and enhances modal discriminability. In $\mathcal{L}_{con}$, a contrastive learning strategy is adopted. By bringing closer and pushing apart the similarity distributions of positive and negative samples, the cross-modal invariant semantic consistency is optimized. The text and audio modality-invariant features of sample i are denoted as I_i^t and I_j^a, respectively. The temperature coefficient τ controls the smoothness of the softmax distribution. After normalization, a cross-modal similarity matrix is constructed as follows:

$$S_{ij} = \frac{(I_i^t, I_j^a)}{\tau} \tag{13}$$

The main diagonal elements are positive samples, and the off-diagonal elements are negative samples. To prevent numerical instability, we use a stable calculation method to model the InfoNCE loss van den Oord et al. (2018):

$$\mathcal{L}_i^{pos} = -\log\left(\frac{\exp(S_{ij})}{\sum_{j=1}^{B}\exp(S_{ij})}\right) \tag{14}$$

In addition, the Top-K hard negative sample mining strategy Huang et al. (2024b) is utilized to select the K most confusing negative instances from the negative samples, forming a strengthened constraint:

$$\mathcal{L}_i^{neg} = \frac{1}{K}\sum_{j\in N_i}\log(1+\exp(S_{ij})) \tag{15}$$

where, N_i represents the set of hard negative samples for sample i. The final loss is comprehensively defined through the similarity weights perceived from labels:

$$\mathcal{L}_{con} = \frac{1}{B}\sum_{i=1}^{B}\omega_i \cdot (\mathcal{L}_i^{pos} + \lambda\mathcal{L}_i^{neg}) \tag{16}$$

where $\lambda = \tau \cdot \sigma(\bar{S}_{\text{pos}})$ is an adaptive temperature adjustment factor based on the average similarity of positive samples. $\sigma(\cdot)$ is the Sigmoid activation function, and $\bar{S}_{\text{pos}}$ is the average similarity score of positive sample pairs.

In $\mathcal{L}_{rep}$, to enhance the discriminative ability of modality-specific features in the semantic space, we design a Triplet-style repulsive loss function. It encourages the specific representations of text and audio modalities to be far away from each other, maintaining the independence of modalities. The text and audio modality-specific features are defined as S_i^t and S_i^a. Similar to constructing the cross-modal similarity matrix, the similarity matrix is defined as follows:

$$R_{ij} = \frac{(S_i^t, S_j^a)}{\tau} \tag{17}$$

To focus on the most interfering negative samples, we select the maximum negative sample similarity term from the off-diagonal elements as $r_i^{\text{hard}} = \max_{i\neq j} R_{ij}$. The modality repulsive loss is then defined as:

$$\mathcal{L}_{rep} = \frac{1}{\sum_i \omega_i}\sum_{i=1}^{B}\omega_i \cdot \max(0, r_i^{hard} - m) \tag{18}$$

where m is the repulsive margin, which is used to control the minimum separation distance. This loss function encourages modality-specific features to be far away from each other in the semantic space, thereby improving modality discriminability and reducing information conflicts.

3.8 MULTI-LABEL CONTRASTIVE LOSS

Building on the preceding context, to fully explore the complex relationships among the semantics of sub-modal labels, we further construct an intra-label aware multi-label loss function $\mathcal{L}_{ml}$, which can be used for feature contrastive learning in this scenario. This loss is based on the overlap measurement between label sets, dynamically adjusting the similarity penalty intensity for different sample pairs. It can both enhance the aggregation of samples with similar semantics and suppress the confusion risk of samples with separated semantics. Thus, by concatenating the prediction results y_i^t and y_i^a for text and audio separately in the model, the multi-label matrix is obtained as $\mathbf{Y} = [\mathbf{y}_1^t, \mathbf{y}_2^t, \ldots, \mathbf{y}_{2B-1}^a, \mathbf{y}_{2B}^a] \in [-1, 1]$, where C is the number of modalities, such as Two modalities are considered semantically related when they have at common label, that is:

$$M_{ij}^{pos} = \mathbf{1}\left[(y_i \cdot y_j^T) > 0\right], \text{s.t. } i \neq j \tag{19}$$

To characterize the semantic similarity degree between samples, we design a label weight matrix $\mathbf{W} \in \mathbb{R}^{B \times B}$. Comprehensively considering the intersection degree $K_s(i,j) = \frac{|y_i \cap y_j|}{|y_i| + \epsilon}$ and the set difference $K_d(i,j) = \frac{1}{1+|y_i| - \|y_i \cap y_j\|}$, the final weight is:

$$W_{ij} = K_s(i,j) \cdot K_d(i,j) \tag{20}$$

This design encourages sub-modal label pairs with greater semantic overlap to obtain higher contrast weights. For this purpose, the construction based on weights is as follows:

$$\mathcal{L}_{ml} = \frac{1}{B} \sum_{i=1}^{B} \sum_{\substack{j \\ j \neq i}} W_{ij} \cdot M_{ij}^{pos} \cdot \left(-\log \frac{\exp(S_{ij})}{\sum_{\substack{k \\ k \neq i}} \exp(S_{ik})} \right) \tag{21}$$

where S_{ij} is the similarity matrix. This loss function can effectively enhance the structural representation ability of sample features under multi-label distribution and avoid misalignment between samples with inconsistent semantics.

Finally, the loss function of the entire model is shown as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{con} + \lambda_{rep}\mathcal{L}_{rep} + \lambda_{ml}\mathcal{L}_{ml} \tag{22}$$

where $\mathcal{L}_{cls}$ is the L1 loss function, and λ_{rep} and λ_{ml} control the balance of the three mechanisms. This joint objective realizes the synergistic learning of the alignment of modality-invariant features and the repulsion of modality-specific features.

4 EXPERIMENTAL

4.1 DATASETS

We conduct experiments on three public multimodal sentiment analysis datasets: CMU-MOSI Zadeh et al. (2016), CMU-MOSEI Bagher Zadeh et al. (2018) and CH-SIMS Yu et al. (2020).

CMU-MOSI consists of 2,199 short monologue video clips from 93 YouTube videos. Each utterance is annotated with a sentiment score from -3 to +3. The dataset is split into 1,284 training, 229 validation, and 686 test samples.

CMU-MOSEI extends CMU-MOSI to 5,000 YouTube videos from over 1,000 speakers and 250 topics, with the same sentiment score range of -3 to +3. It includes 16,326 training, 1,871 validation, and 4,659 test samples.

CH-SIMS is a Chinese dataset with 2,281 video clips and sub-modal sentiment annotations from -1 to +1. The dataset is divided into 1,368 training, 456 validation, and 457 test samples.

4.2 EVALUATION METRICS

Following previous work Hazarika et al. (2020b), we used the accuracy of 2-class (Acc2) and 5-class (Acc5) on CH-SIMS, the accuracy of 2-class (Acc2) on MOSI and MOSEI. Mean Absolute Error (MAE), Pearson Correlation (Corr), and F1-score (F1) on all datasets. Moreover, on MOSI and MOSEI, Acc2 and F1 used two calculation ways: negative/non-negative (has-0) and negative/positive (non-0). Except for MAE, higher values indicate better performance for all metrics.

4.3 Implementation Details

We implement the proposed model using the PyTorch framework and conduct all training on a single NVIDIA L20 GPU. For audio preprocessing, we sample audio signals at 16 kHz with a fixed duration of 6 seconds. Audio segments longer than 6 seconds are truncated, while shorter segments are zero-padded at the end to ensure all audio features are standardized to 96,000 sampling points. For model optimization, we employ the AdamW optimizer with a warm-up strategy and cosine annealing learning rate scheduler. The hyperparameters are dataset-specific: for CMU-MOSI and CMU-MOSEI datasets, we set the learning rate to 5e-6 with a batch size of 8, while for the CH-SIMS dataset, we use a learning rate of 1e-5 with a batch size of 16.

4.4 Baseline Multimodal models

To comprehensively evaluate the effectiveness of the proposed SLaD model, we compare it with a diverse set of existing MSA approaches, including both representation-based and advanced interaction-based methods. Representation-based methods include: MISA Hazarika et al. (2020a), MMM Han et al. (2021), Self-MM Yu et al. (2021) and FMFN Li et al. (2025a). Interaction-based methods include: TFN Zadeh et al. (2017), LMF Liu et al. (2018), MulT Tsai et al. (2019), CENet Wang et al. (2023b), TETFN Wang et al. (2023a), ALMT Zhang et al. (2023), TMBL Huang et al. (2024a) and KuDA Feng et al. (2024).

Table 1: Model comparison results on the CMU-MOSI and CMU-MOSEI datasets. "^{a}" denotes results reproduced using the authors' released code, "^{b}" indicates results reported in Li et al. (2025a), and "^{c}" indicates results reported in the original paper.

Model	CMU-MOSI				CMU-MOSEI			
	Acc2	F1	MAE	Corr	Acc2	F1	MAE	Corr
TFN^a	76.9/78.2	76.9/78.2	0.962	0.658	82.0/82.7	81.9/82.2	0.572	0.718
MISA^a	82.5/83.8	82.5/83.9	0.757	0.787	81.2/84.7	81.7/84.7	0.544	0.763
MMM^b	83.6/85.3	83.6/85.3	0.755	0.773	83.2/85.0	83.4/84.8	0.543	0.758
CENet^a	83.5/85.2	83.4/85.2	0.725	0.795	83.5/86.3	83.8/86.3	0.525	0.777
TETEN^b	84.0/86.1	83.8/86.0	0.717	0.800	84.2/85.1	84.1/85.2	0.551	0.748
ALMT^a	84.5/86.4	84.5/86.4	<u>0.683</u>	<u>0.805</u>	<u>84.7/86.7</u>	<u>85.1/86.8</u>	**0.526**	<u>0.779</u>
KuDA^c	84.4/86.4	84.5/86.5	0.705	0.794	83.3/86.5	83.0/86.6	<u>0.529</u>	0.776
FMFH^b	<u>84.8/87.0</u>	<u>85.0/87.1</u>	0.728	0.794	83.2/86.3	82.8/86.4	0.533	0.774
SLaD(our)	**87.8/90.2**	**87.7/90.2**	**0.596**	**0.874**	**86.2/88.2**	**86.3/88.0**	0.555	**0.812**

4.5 Comparison of Results

Table 1 and Table 2 present comprehensive comparative results of our proposed SLaD method against baseline models on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets.

4.5.1 Performance on CMU-MOSI and CMU-MOSEI Datasets

As demonstrated in Table 1, SLaD achieves superior performance across all evaluation metrics on both datasets. On the CMU-MOSI dataset, our model demonstrates substantial improvements over the second-best baseline FMFH: a remarkable 3.0%/3.2% improvement in Acc2, a 2.7%/3.1% enhancement in F1-score, a significant reduction of 0.087 in MAE, and a notable 0.069 increase in correlation coefficient. These improvements indicate SLaD's enhanced capability in sentiment polarity classification and regression tasks.

Similarly, on the CMU-MOSEI dataset, SLaD consistently outperforms all baseline methods. Compared to the second-best performing ALMT, our model achieves a 1.5%/1.5% improvement in Acc2, a 1.2%/1.2% enhancement in F1-score, and a substantial 0.033 increase in correlation coefficient. While showing a slight increase of 0.029 in MAE compared to ALMT, the overall performance gains across other metrics demonstrate the robustness and effectiveness of our approach.

Table 2: Model comparison results on the CH-SIMS dataset. "[a]" denotes results reproduced using the authors' released code, "[b]" indicates results reported in Li et al. (2025a), and "[c]" indicates results reported in the original paper.

Model	Acc2	F1	Acc5	MAE	Corr
TFN[a]	77.8	78.0	36.3	0.442	0.575
LMF[a]	77.2	77.4	41.0	0.439	0.586
MulT[a]	76.2	76.4	36.5	0.441	0.588
Self-MM[a]	78.9	78.8	<u>44.8</u>	0.410	0.600
CENet[b]	77.9	77.5	33.9	0.470	0.539
ALMT[b]	79.4	79.6	42.4	0.420	0.594
TMBL[b]	79.1	78.7	41.5	0.429	0.592
KuDA[c]	<u>80.7</u>	<u>80.7</u>	43.5	<u>0.408</u>	<u>0.613</u>
FMFN[b]	<u>80.7</u>	<u>80.7</u>	44.2	0.416	0.598
SLaD(our)	**81.6**	**81.4**	**47.9**	**0.390**	**0.631**

4.5.2 Performance on CH-SIMS Dataset

The CH-SIMS dataset presents more complex multimodal scenarios with richer contextual information, making it particularly challenging for MSA tasks. As shown in Table 2, SLaD achieves state-of-the-art performance across all evaluation metrics, demonstrating its superior capability in handling complex multimodal sentiment analysis scenarios.

Specifically, SLaD outperforms the best baseline models KuDA and FMFN by 0.9% in both Acc2 and F1-score for binary classification tasks. In the more challenging five-class classification task (Acc5), our method achieves a substantial 3.1% improvement over the second-best baseline Self-MM, reaching 47.9% accuracy. For regression tasks, SLaD reduces MAE by 0.018 compared to the best baseline KuDA and increases the correlation coefficient by 0.018, indicating superior performance in fine-grained sentiment intensity prediction.

4.6 Ablation Study and Analysis

To systematically validate the effectiveness of each key component, comprehensive ablation experiments were conducted on the CH-SIMS dataset. By selectively removing or combining different mechanisms, including the sub-modality label-aware similarity weighting ω, modality contrastive loss $\mathcal{L}_{con}$, modality repulsion loss $\mathcal{L}_{rep}$, and multi-label supervision loss $\mathcal{L}_{ml}$, we investigated their contributions and synergistic effects in cross-modal emotion prediction. The full results are summarized in Table 3.

4.6.1 Effectiveness of Sub-modality Label-aware Similarity Weighting

Removing similarity weighting ω significantly hurt performance (Acc2, F1, Corr dropped by 4%, MAE increased), showing its importance in aligning cross-modal labels and reducing noise. Without modality contrastive loss ($\mathcal{L}_{con}$), performance worsened further (7% Acc2, 6% F1 drop), highlighting the synergy between ω and $\mathcal{L}_{con}$ for label consistency. Retaining only ω yielded the worst results, indicating its limited independent value. It has maximal utility when integrated with other components.

4.6.2 Effectiveness of Modality Contrastive Loss

The effectiveness of the modality contrastive loss $\mathcal{L}_{con}$ is reflected in approximately 3% reductions in Acc2 and F1 and a marked increase in MAE upon its removal, highlighting its critical function in aligning modality-invariant features across modalities. Further comparison shows that when $\mathcal{L}_{con}$ is used solely alongside ω and $\mathcal{L}_{ml}$, the performance declines further, indicating a complementary relationship between the contrastive loss and repulsion loss in modeling consistency and discriminability. The poor performance when $\mathcal{L}_{con}$ is used alone further confirms that this loss requires the structural semantic constraints provided by other mechanisms to realize its full potential.

Table 3: Ablation study on the CH-SIMS dataset. The percentage in parentheses indicates the relative change compared with the full model ($\omega + \mathcal{L}_{con} + \mathcal{L}_{rep} + \mathcal{L}_{ml}$), calculated as (current − full)/full × 100%.

ω	$\mathcal{L}_{con}$	$\mathcal{L}_{rep}$	$\mathcal{L}_{ml}$	**Acc2**	**F1**	**Acc5**	**MAE**	**Corr**
✓	✓	✓	✓	81.6	81.4	47.9	0.390	0.632
	✓	✓	✓	77.9 (↓4.55%)	78.4 (↓3.69%)	42.9 (↓10.44%)	0.432 (↑10.8%)	0.608 (↓3.80%)
✓		✓	✓	75.5 (↓7.48%)	76.2 (↓6.39%)	46.8 (↓2.29%)	0.400 (↑2.56%)	0.587 (↓7.12%)
✓	✓		✓	78.8 (↓3.43%)	79.0 (↓2.95%)	43.5 (↓9.19%)	0.444 (↑13.8%)	0.604 (↓4.43%)
✓	✓	✓		77.5 (↓5.02%)	78.2 (↓3.93%)	44.0 (↓8.15%)	0.413 (↑5.9%)	0.636 (↑0.63%)
		✓	✓	79.0 (↓3.18%)	78.9 (↓3.07%)	45.1 (↓5.84%)	0.377 (↓3.33%)	0.634 (↑0.32%)
	✓		✓	75.7 (↓7.23%)	76.5 (↓6.03%)	42.9 (↓10.44%)	0.425 (↑9.0%)	0.613 (↓3.01%)
	✓	✓		78.3 (↓4.04%)	78.7 (↓3.32%)	46.4 (↓3.13%)	0.386 (↓1.03%)	0.641 (↑1.42%)
✓			✓	76.2 (↓6.61%)	76.9 (↓5.52%)	46.8 (↓2.29%)	0.392 (↑0.51%)	0.607 (↓3.95%)
✓		✓		74.4 (↓8.82%)	75.2 (↓7.62%)	46.4 (↓3.13%)	0.410 (↑5.13%)	0.577 (↓8.70%)
✓	✓			76.4 (↓6.37%)	77.0 (↓5.40%)	47.9 (=0.00%)	0.383 (↓1.79%)	0.636 (↑0.63%)
			✓	76.2 (↓6.61%)	76.7 (↓5.78%)	48.6 (↑1.46%)	0.401 (↑2.82%)	0.600 (↓5.06%)
	✓			77.9 (↓4.55%)	78.1 (↓4.06%)	45.3 (↓5.43%)	0.408 (↑4.62%)	0.635 (↑0.47%)
		✓		74.2 (↓9.06%)	75.1 (↓7.75%)	46.0 (↓3.96%)	0.388 (↓0.51%)	0.622 (↓1.58%)
				77.4 (↓5.14%)	77.9 (↓4.30%)	43.3 (↓9.61%)	0.409 (↑4.87%)	0.616 (↓2.53%)

4.6.3 Effectiveness of Modality Repulsion Loss

Regarding the modality repulsion loss $\mathcal{L}_{rep}$, its removal resulted in approximately 5% decreases in Acc2 and F1 and nearly a 6% increase in MAE, demonstrating that it effectively enhances the discriminability of modality-specific features and suppresses cross-modal interference. Notably, correlation slightly improved when combined with other components, suggesting that $\mathcal{L}_{rep}$ maintains modality independence while complementing consistency modeling to optimize overall performance.

4.6.4 Effectiveness of Multi-label Supervision Loss

The multi-label supervision loss $\mathcal{L}_{ml}$ exhibited the most significant impact, with its exclusion causing over 6% and 5% reductions in Acc2 and F1, respectively, along with approximately a 5% drop in correlation. This emphasizes its indispensable role in mining semantic relationships among labels and optimizing cross-modal information fusion. The substantial performance degradation observed when $\mathcal{L}_{ml}$ is retained alone further indicates its limited independent efficacy, making it more suitable for collaborative use with other mechanisms.

In summary, these four mechanisms mutually complement and collaboratively optimize the model at different levels, collectively enhancing cross-modal emotion recognition performance.

5 Conclusion

In this work, we propose SLaD, a novel framework designed to disentangle modality-invariant and modality-specific representations for MSA. By leveraging a sub-modal label similarity weighting mechanism, SLaD effectively models the structural and numerical relationships among sub-modal emotional labels, ensuring consistent and reliable cross-modal alignment. Moreover, through the integration of a modality contrastive loss, modality repulsive loss, and multi-label contrastive loss, SLaD achieves both semantic consistency and enhanced modality discriminability. Experimental results on three benchmark datasets demonstrate that SLaD achieves SOTA performance, significantly improving both classification and regression metrics. In future work, we plan to extend SLaD to incorporate visual modalities and explore cross-domain adaptation further to enhance its generalization capability in diverse real-world applications.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika

Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208.

Chen, Sufen, Liu, Ye, Zeng, and Xueqiang. Fusing emoji emotion distribution for multi-label emotion classification. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pp. 129–133, 2024.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 657–668, Online, November 2020. Association for Computational Linguistics.

Deng, Jiawen, Ren, and Fuji. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486, 2023.

Cunhang Fan, Kang Zhu, Jianhua Tao, Guofeng Yi, Jun Xue, and Zhao Lv. Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 16(1):207–222, 2025.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14755–14766, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Chenquan Gan, Yu Tang, Xiang Fu, Qingyi Zhu, Deepak Kumar Jain, and Salvador García. Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. *Knowledge-Based Systems*, 299:111982, 2024. ISSN 0950-7051.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 1122–1131, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450379885.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *CoRR*, abs/2005.03545, 2020b. URL https://arxiv.org/abs/2005.03545.

Jiehui Huang, Jun Zhou, Zhenchao Tang, Jiaying Lin, and Calvin Yu-Chian Chen. Tmbl: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems*, 285:111346, 2024a. ISSN 0950-7051.

Wentao Huang, Xiaoling Hu, Shahira Abousamra, Prateek Prasanna, and Chao Chen. Hard Negative Sample Mining for Whole Slide Image Classification . In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15004. Springer Nature Switzerland, October 2024b.

Weiqiang Jin, Biao Zhao, Yu Zhang, Jia Huang, and Hang Yu. Wordtransabsa: Enhancing aspect-based sentiment analysis with masked language modeling for affective token prediction. *Expert Systems with Applications*, 238:122289, 2024. ISSN 0957-4174.

Kyeonghun Kim and Sanghyun Park. Aobert: All-modalities-in-one bert for multimodal sentiment analysis. *Information Fusion*, 92:37–45, 2023. ISSN 1566-2535.

Kai Li, Cheng Zhou, Xin (Robert) Luo, Jose Benitez, and Qinyu Liao. Impact of information timeliness and richness on public engagement on social media during covid-19 pandemic: An empirical investigation based on nlp and machine learning. *Decision Support Systems*, 162:113752, 2022. ISSN 0167-9236. Business and Government Applications of Text Mining & Natural Language Processing (NLP) for Societal Benefit.

Xiang Li, Haijun Zhang, Zhiqiang Dong, Xianfu Cheng, Yun Liu, and Xiaoming Zhang. Learning fine-grained representation with token-level alignment for multimodal sentiment analysis. *Expert Systems with Applications*, 269:126274, 2025a. ISSN 0957-4174.

Xingye Li, Jin Liu, Yurong Xie, Peizhu Gong, Xiliang Zhang, and Huihua He. Magdra: A multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition. *Knowledge-Based Systems*, 283:111126, 2024. ISSN 0950-7051.

Zuhe Li, Qingbing Guo, Yushan Pan, Weiping Ding, Jun Yu, Yazhou Zhang, Weihua Liu, Haoran Chen, Hao Wang, and Ying Xie. Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis. *Information Fusion*, 99:101891, 2023. ISSN 1566-2535.

Zuhe Li, Panbo Liu, Yushan Pan, Weiping Ding, Jun Yu, Haoran Chen, Weihua Liu, Yiming Luo, and Hao Wang. Multimodal sentiment analysis based on disentangled representation learning and cross-modal-context association mining. *Neurocomputing*, 617:128940, 2025b. ISSN 0925-2312.

Ronghao Lin and Haifeng Hu. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(2):549–565, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Nusrat J. Shoumy, Li-Minn Ang, Kah Phooi Seng, D.M.Motiur Rahaman, and Tanveer Zia. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447, 2020. ISSN 1084-8045.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL `http://arxiv.org/abs/1807.03748`.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259, 2023a. ISSN 0031-3203.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023b.

Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, and Li'an Zhu. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management*, 61(3):103675, 2024. ISSN 0306-4573.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. Multimodal multi-loss fusion network for sentiment analysis. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3588–3602, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508, 2023. ISSN 0306-4573.

Lixun Xie, Weiqing Sun, Jingyi Zhang, and Xiaohu Zhao. Ac2net: Hybrid attention convolution and compression fusion network for multimodal emotion recognition. *Digital Signal Processing*, 164:105261, 2025. ISSN 1051-2004.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. Multiple contrastive learning for multimodal sentiment analysis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Online, July 2020. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797, May 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016. URL http://arxiv.org/abs/1606.06259.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics.

Shuting Zheng, Jingling Zhang, Yuanzhao Deng, and Lanxiang Chen. Cmff: A cross-modal multi-layer feature fusion network for multimodal sentiment analysis. *Applied Soft Computing*, 184: 113868, 2025. ISSN 1568-4946.

Chuanbo Zhu, Min Chen, Sheng Zhang, Chao Sun, Han Liang, Yifan Liu, and Jincai Chen. Skeafn: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. *Information Fusion*, 100:101958, 2023. ISSN 1566-2535.

Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, Ken Zheng, and Qunyan Zhou. Acformer: An aligned and compact transformer for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 833–842, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085.

Roman Zubatyuk, Justin S. Smith, Jerzy Leszczynski, and Olexandr Isayev. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science Advances*, 5(8):eaav6490, 2019.

A APPENDIX

A.1 REPRODUCIBILITY CHECKLIST

All experiments are implemented using the PyTorch 2.5.1 framework. The training is conducted on a platform equipped with an Intel (R) Xeon (R) Platinum 8457C CPU and an NVIDIA L20 GPU. The key hyper-parameters are summarized in Table 4. Empirical results show that the optimal settings of most hyperparameters remain consistent across all datasets, highlighting the robustness of SLaD with respect to hyperparameter selection. This property alleviates the need for labor-intensive dataset-specific hyperparameter tuning.

Table 4: Hyper-parameters of SLaD we use on the different datasets.

	CMU-MOSI	CMU-MOSEI	CH-SIMS
Learning Rate	5e-6	5e-6	1e-5
Batch Size	8	8	16
Optimizer	AdamW	AdamW	AdamW
Epochs	50	50	50
Warm Up	✓	✓	✓
Cosine Annealing	✓	✓	✓
Context	✓	✓	-
Text Context Length	2	2	-
Audio Context Length	2	2	-
Early Stop	5	5	5
Dropout	0.3	0.3	0.3
Bandwidth Parameter α	0.5	0.5	0.5
Temperature Coefficient τ	0.2	0.2	0.2
Top-K	3	3	3
Number Attention Heads	12	12	12

A.2 BIMODAL EXPERIMENT

In this section, we compare the proposed method with other state-of-the-art bimodal models.

A.2.1 BASELINE BIMODAL MODELS

Self-MM Yu et al. (2021): This model generates sub-modality labels and further introduces a weight adjustment strategy to balance the learning progress across different sub-tasks.

AOBERT Kim & Park (2023): This model jointly trains a multimodal masked language model and an alignment prediction task to identify dependencies and correlations among modalities.

SKEAFN Zhu et al. (2023): This framework incorporates external affective knowledge representations to enhance the textual modality. It employs a text-guided interaction module to promote interactions between text and other modalities, while a feature-level attention mechanism dynamically adjusts weights during multimodal fusion.

MCMF Li et al. (2023): This model proposes a linguistically guided Transformer framework with unimodal feature fusion to address the challenges of multimodal information integration.

MTMD Lin & Hu (2024): In this model, textual and multimodal representations are regarded as teacher networks, while acoustic and visual representations serve as student networks, enabling knowledge distillation. The distillation process leverages both regression and classification subtasks to facilitate the transfer of sentiment-related knowledge.

MMML Wu et al. (2024): This study shows that pre-trained models for raw audio enhance feature extraction, and that combining audio with text outperforms using text alone. Moreover, multi-loss training and contextual information substantially improve multimodal sentiment analysis.

CMFF Zheng et al. (2025): CMFF exploits hierarchical information embedded in shallow and deep features of text and audio. A multi-head cross-modal attention mechanism is employed in the fusion layer to facilitate interactions across feature levels and modalities.

Table 5: Model comparison results on the CMU-MOSI and CMU-MOSEI datasets, "*" indicates results reported in Zheng et al. (2025).

Model	CMU-MOSI				CMU-MOSEI			
	Acc2	F1	MAE	Corr	Acc2	F1	MAE	Corr
Self-MM*	84.0/86.0	84.4/86.0	0.713	0.798	82.8/85.2	82.5/85.3	0.530	0.765
AOBERT*	85.2/85.6	85.4/86.4	0.856	0.700	84.9/86.2	85.0/85.9	0.515	0.763
SKEAFN*	85.1/87.3	85.2/87.3	0.665	0.825	84.3/87.1	84.1/87.2	0.517	0.788
MCMF*	85.2/88.4	85.3/88.4	0.690	0.810	84.7/86.2	84.7/85.9	<u>0.510</u>	0.740
MTMD*	84.0/86.0	83.9/86.0	0.705	0.799	84.8/86.1	84.9/85.9	0.531	0.767
MMML*	85.9/88.2	85.9/88.2	0.643	0.838	86.3/86.7	86.2/86.5	0.517	0.791
CMFF*	<u>87.3/89.2</u>	<u>87.3/89.2</u>	**0.570**	**0.876**	<u>84.0/88.2</u>	<u>84.4/88.2</u>	**0.483**	**0.813**
SLaD(our)	**87.8/90.2**	**87.7/90.2**	<u>0.596</u>	<u>0.874</u>	**86.2/88.2**	**86.3/88.0**	0.555	<u>0.812</u>

A.2.2 Comparison of Results

Table 5 reports results on CMU-MOSI and CMU-MOSEI under the bimodal (text + audio) setting. On CMU-MOSI, SLaD achieves state-of-the-art classification accuracy, outperforming CMFF by up to +1.0% (Acc2/F1) and yielding consistent gains over MMML and SKEAFN, which confirms the effectiveness of sub-modal label-aware supervision in capturing fine-grained polarity. For regression, SLaD attains Corr nearly identical to CMFF (0.874 vs. 0.876) with a slightly higher MAE, while showing clear improvements over other baselines. On CMU-MOSEI, SLaD remains highly competitive: it provides balanced improvements over MMML (+1.5% Acc2/F1 under the negative–positive paradigm) and stronger robustness than CMFF in the non-negative setting (+2.2%/+1.9% Acc2/F1). Although its MAE is higher, Corr remains comparable to CMFF (0.812 vs. 0.813) and superior to most baselines (e.g., +0.024 Corr relative to SKEAFN). Overall, SLaD consistently advances classification while maintaining competitive regression performance, demonstrating robustness across datasets.

A.3 Influence of Temperature Coefficient and Hard Negative Mining Strategy

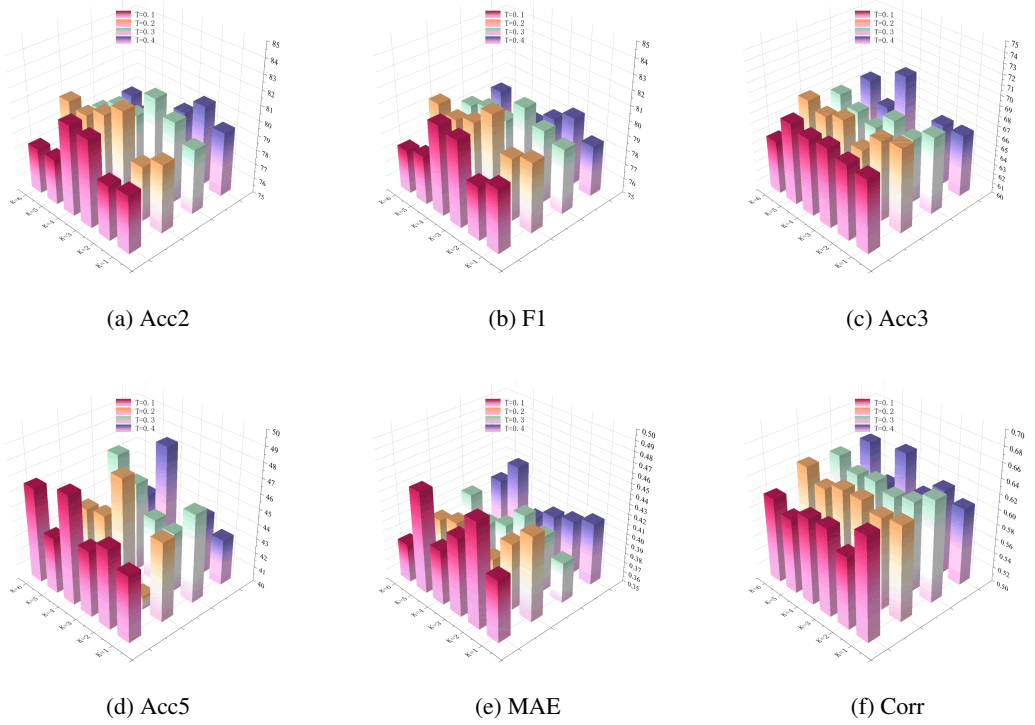
To verify the effectiveness of the temperature coefficient τ and the Top-K hard negative sample mining strategy in the proposed modal contrastive loss $\mathcal{L}_{con}$, we conducted systematic experiments under different temperature coefficients ($\tau = 0.1, 0.2, 0.3, 0.4$) and different numbers of negative samples ($K = 1, 2, 3, 4, 5, 6$). Figure 4 illustrates the influence of the temperature coefficient and hard negative mining strategy.

A.3.1 Influence of Temperature Coefficient

The temperature coefficient τ in $\mathcal{L}_{con}$ has a critical impact on cross-modal alignment. A small value ($\tau = 0.1$) sharpens the similarity distribution excessively, causing unstable optimization and noisy hard negative selection, with F1 fluctuating between 77.9%–81.1% (e.g., 78.3% at ($K = 5$)) and Corr around 0.615. In contrast, moderate values ($\tau = 0.2 \sim 0.3$) balance gradients across pairs and improve hard negative mining, achieving the best performance at ($\tau = 0.2, K = 3$) with F1=81.4%, MAE=0.390, and Corr=0.631. Larger values ($\tau = 0.4$) overly smooth the distribution, reducing discriminability: although Corr peaks at 0.640 ($K = 4$), F1 drops to 77.0%–78.9%. These results indicate that $\tau = 0.2$ offers the best trade-off between stability and discriminability.

A.3.2 Influence of Hard Negative Mining Strategy

The Top-K hard negative mining strategy is introduced to enhance feature discriminability by prioritizing the most confusing negatives. Experimental results across different K values validate this



(a) Acc2 (b) F1 (c) Acc3

(d) Acc5 (e) MAE (f) Corr

Figure 4: Impact of different temperature coefficient τ and Top-K hard negative mining strategy K.

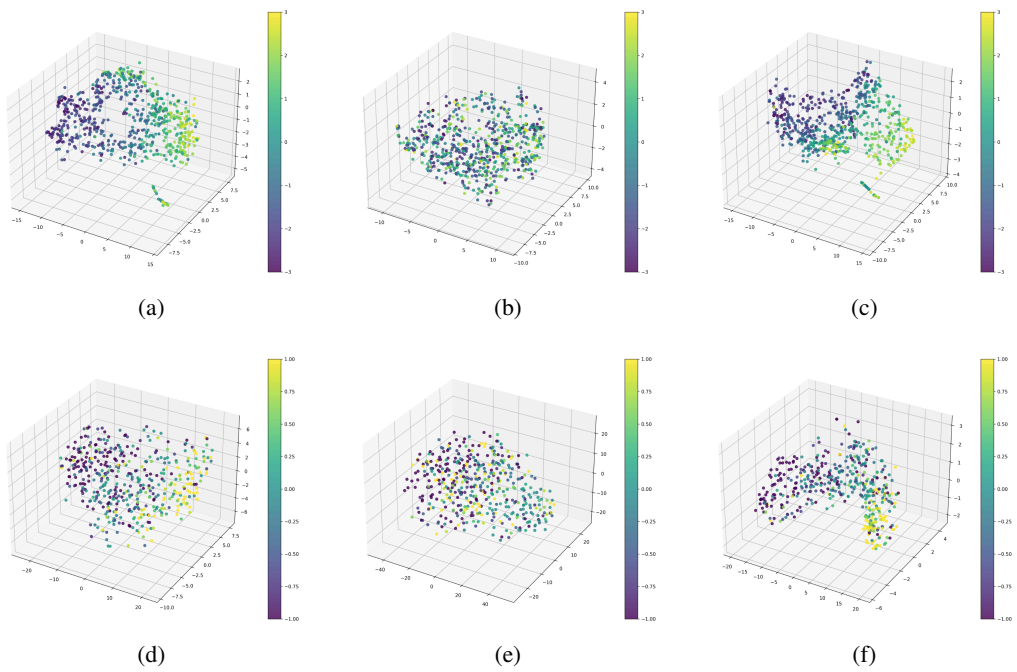
design. Small K (e.g., $K = 1, 2$) imposes limited constraints, failing to fully exploit semantically confusing negatives and thus yielding suboptimal performance. For example, at $\tau = 0.2$, F1 reaches only 79.7% for $K = 1$, compared to 81.4% for $K = 3$. Moderate values ($K = 3$) best capture the hardest negatives, achieving optimal performance at ($\tau = 0.2, K = 3$) with F1=81.4%, MAE=0.390, and Corr=0.631.

By contrast, large K (e.g., $K \geq 5$) dilutes the hard negative set with numerous easily distinguishable negatives, which provide little gradient signal and instead introduce noise. For instance, at ($\tau = 0.1, K = 5$), performance drops notably (F1=78.3%, Corr=0.600).

Overall, a moderate temperature coefficient and a carefully chosen K are both critical for cross-modal consistency modeling. The temperature balances gradients between positive and negative pairs, while the Top-K strategy sharpens modality discrimination by emphasizing the most confusing negatives. Together, these mechanisms jointly improve alignment and discriminability in multimodal sentiment analysis.

A.4 Visualization of the input unimodal features and final fused features

As shown in Figure 5, for the binary sentiment classification task, we employ the T-SNE method to visualize the input unimodal features (i.e., X'^t and X'^a) and the final fused representation X^f on the CMU-MOSI and CH-SIMS datasets. The results reveal that, compared to unimodal features (whose sample distributions exhibit substantial overlap), the fused representation demonstrates markedly stronger separability, with clear boundaries emerging between positive and negative sentiment classes. This observation indicates that SLaD effectively leverages the diverse and complementary information across modalities, thereby providing robust support for improving sentiment classification accuracy.



(a) (b) (c)

(d) (e) (f)

Figure 5: Visualization of the input unimodal features and final fused features on MOSI and SIMS datasets. (a) Input Text Features of MOSI, (b) Input Audio Features of MOSI, (c) Final Fused Features of MOSI, (d) Input Text Features of SIMS, (e) Input Audio Features of SIMS, (f)Final Fused Features of SIMS.

A.5 METRICS

We adopt both classification and regression metrics for evaluation. Note that sentiment intensity ranges from [-1,1] for CH-SIMS and [-3,3] for CMU-MOSI and CMU-MOSEI.

Classification metrics:

- **Acc2**: Binary accuracy (positive vs. negative, or non-negative vs. negative).
- **Acc5**: Accuracy under a five-class setting by discretizing the sentiment range into five intervals.
- **F1**: Harmonic mean of precision and recall, weighted across classes to mitigate class imbalance.

Regression metrics:

- **MAE**: Mean absolute error between predicted and ground-truth scores.
- **Corr**: Pearson correlation coefficient measuring linear correlation between predictions and labels.

For the Acc2 and F1 score, we describe two cases as follows:

- **Non-negative/Negative (NN/N)**: This classification is used to evaluate the model's ability to distinguish between non-negative (greater than or equal to 0) and negative (less than 0) sentiments.
- **Positive/Negative (P/N)**: This configuration focuses on the accuracy of the model in classifying sentiment as positive (greater than 0) or negative (less than 0).

A.6 LIMITATION

A key limitation of this work lies in its current focus on text–audio bimodality, without incorporating the visual modality. This design choice is motivated by three considerations: (i) prior studies have shown that adding visual input often brings marginal performance gains in related tasks Wu et al. (2024); Zheng et al. (2025); (ii) visual integration substantially increases computational cost and hardware requirements; and (iii) visual information is frequently unavailable in real-world applications, limiting the practicality of models that rely on it.

Despite this restriction, the proposed model achieves state-of-the-art performance across both bimodal and multimodal task settings. Nevertheless, extending the framework to include the visual modality remains an important future direction. Enhancing the model's visual understanding capability may further improve its effectiveness in comprehensive multimodal scenarios.

A.7 THE USE OF LARGE LANGUAGE MODELS

During the preparation of this manuscript, Large Language Models (LLMs) were employed as a general-purpose writing assistant. Their role was limited to grammar correction, style refinement, and improving the clarity of exposition. Importantly, LLMs were not used for research ideation, methodological design, experimental setup, data analysis, or interpretation of findings. All scientific content, results, and conclusions presented in this paper are solely the responsibility of the authors.