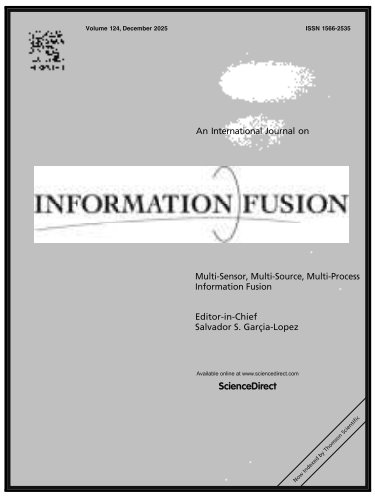

Divide-and-Conquer: Prompt-based Distribution Learning for Multimodal Sentiment Analysis

Chengji Wang, Wenjing Zhang, Guanyi Chen, Ming Dong, Tingting He, Xingpeng Jiang

PII: S1566-2535(26)00172-7
DOI: https://doi.org/10.1016/j.inffus.2026.104293
Reference: INFFUS 104293

To appear in: *Information Fusion*

Received date: 30 May 2025
Revised date: 23 January 2026
Accepted date: 9 March 2026

Please cite this article as: Chengji Wang, Wenjing Zhang, Guanyi Chen, Ming Dong, Tingting He, Xingpeng Jiang, Divide-and-Conquer: Prompt-based Distribution Learning for Multimodal Sentiment Analysis, *Information Fusion* (2025), doi: https://doi.org/10.1016/j.inffus.2026.104293

Highlights

- Ambiguity in multimodal signals can be represented as a learnable distribution.
- We transform the sentiment prediction task into a distribution learning problem.
- We integrate prompt tuning with feature decomposition to mine sentiment distributions.
- We penalize incorrect proximities between distributions and encourage correct ones.
- We propose a relationship-augmented feature fusion module.
- Experimental results validate the effectiveness of the proposed framework.

Divide-and-Conquer: Prompt-based Distribution Learning for Multimodal Sentiment Analysis

Chengji Wang[a,b,c], Wenjing Zhang[a,b,c], Guanyi Chen[a,b,c], Ming Dong[a,b,c], Tingting He[a,b,c,*], Xingpeng Jiang[a,b,c,*]

[a]*Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, 430079, Hubei, China*
[b]*National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, Wuhan, 430079, Hubei, China*
[c]*School of Computer Science, Central China Normal University, Wuhan, 430079, Hubei, China*

Abstract

While multimodal fusion with injective embedding has shown significant effectiveness in multimodal sentiment analysis, it is constrained by the requirement of a single-point embedding and the intrinsic ambiguity within multimodal signals, limiting its practical applications. In this paper, we propose Prompt-based Distribution Learning (PDL), a novel divide-and-conquer framework designed for multimodal sentiment analysis. PDL aims to enhance the representation of multimodal information by jointly performing prompt tuning and sentiment distribution learning. Specifically, PDL utilizes modality-aware prompts as probes to decompose multimodal features and mine the latent sentiment distributions for each modality, thereby capturing a wide range of sentiment information. We conduct extensive experiments on three benchmark datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS, to evaluate PDL. The experimental results validate the efficacy of PDL, showing that PDL outperforms existing approaches in capturing and representing the complex sentiment information present in multimodal data.

Keywords: Multimodal Sentiment Analysis, Multimodal Fusion, Prompt Tuning, Label Distribution Learning

1. Introduction

Multimodal Sentiment Analysis (MSA) aims to leverage multiple data modalities, such as text, audio and visual, to predict sentimental polarity [1]. Most of MSA methods [2–4] focus on learning an injective embedding to convert multimodal data into numeric vectors. This approach implicitly frames MSA as a

*Corresponding authors: Tingting He, Xingpeng Jiang
Email addresses: wcj@ccnu.edu.cn (Chengji Wang), dongri_z@mails.ccnu.edu.cn (Wenjing Zhang), g.chen@ccnu.edu.cn (Guanyi Chen), dongming@ccnu.edu.cn (Ming Dong), tthe@ccnu.edu.cn (Tingting He), xpjiang@ccnu.edu.cn (Xingpeng Jiang)

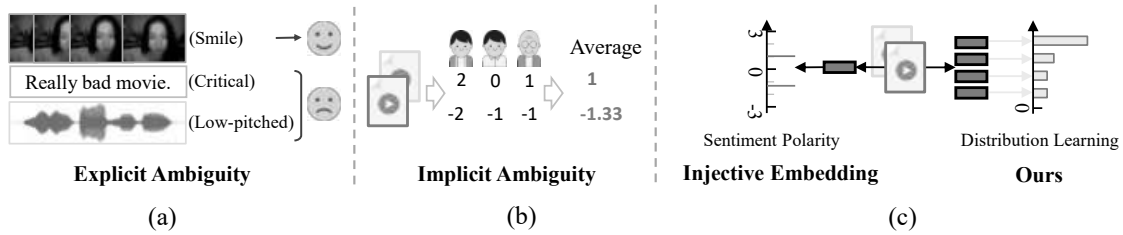

Figure 1: Explicit and implicit ambiguity in multimodal sentiment analysis. We propose to learn a sentiment distribution through a one-to-many mapping, rather than obtaining an injective embedding with an ambiguous label.

single-label learning problem, assuming that multimodal data possesses a single and clearly defined sentiment label.

Compelling evidence [2, 3, 5–7] challenges the practical validity of assuming clear, consistent multimodal signals, revealing it as an idealization. Inherent ambiguities arising from redundancy and noise not only adversely affect model performance but also prevent the formation of an injective mapping in the embedding space. Data ambiguities can be classified into two types: explicit and implicit. Explicit ambiguity arises when multimodal data simultaneously convey conflicting sentiments. For example, in Figure 1(a), a smiling face suggests a positive sentiment, whereas the spoken word "bad" and a low-pitched voice imply a negative one. This conflict makes definitive sentiment labeling challenging [3, 5, 8]. Implicit ambiguity refers to labeling inconsistencies caused by annotator disagreement. As in Figure 1(b), disagreements may arise from subjective biases or cultural variations, which complicates the attainment of a consistent label [2, 9].

In response, recent studies [7, 10, 11] have proposed gaussian embedding networks to capture these data ambiguities via probabilistic embeddings. Instead of producing a deterministic point estimate, this method parameterizes a gaussian distribution by optimizing its mean and covariance vectors, with the final representation sampled from this distribution. Although these approaches have yielded promising results, a gaussian distribution is insufficient to adequately address both implicit and explicit ambiguities. Crucially, we argue that the objective should be to faithfully simulate the human labeling process, rather than to train a model that merely passively fits the distribution of an existing, ambiguously labeled dataset.

To address the aforementioned challenges, we propose a novel divide-and-conquer framework named Prompt-based Distribution Learning (PDL), which reformulates sentiment prediction as a distribution learning problem to learn more comprehensive and robust representations against data ambiguities. As shown in Figure 1(c), PDL simulates the human annotation process by predicting a distribution of potential labels. To tackle implicit ambiguity, PDL employs a set of learnable prompts, each representing a distinct annotator. These prompts act as semantic probes to decompose unimodal features into multiple fine-grained embeddings, with each capturing a different facet of sentiment semantics. To handle explicit ambiguity, we define these prompts to be modality-aware. We further introduce a Text-Conditioned Feature Decomposition (TCFD) module, which leverages these modality-aware prompts to guide the decomposition of multimodal features, thereby constructing separate sentiment distributions for visual and audio modalities. A Contrastive Distribution Learning method (CDL) is then introduced to align the learned multimodal sentimental distributions, enforcing that instances with the same sentiment label have similar sentiment distributions. We also propose a Relation-Augmented Feature Fusion (RAFF) module, which treats the learned sentiment distributions as importance weights to dynamically fuse the fine-grained embeddings, thereby reinforcing salient sentiment cues across modalities. By integrating these components, PDL adaptively addresses both explicit and implicit ambiguities, resulting in a more robust and interpretable model for real-world multimodal sentiment analysis.

For evaluation, PDL was verified on three multimodal sentiment benchmarks, CMU-MOSI [12], CMU-MOSEI [13] and CH-SIMS [4]. The experimental results show that PDL achieves state-of-the-art or even superior performance compared to baseline models. Additionally, compared to baseline models, our framework consistently improves performance by effectively combining sentiment-related features from individual modalities. This approach maximizes the unique contributions of each modality, enhancing the overall

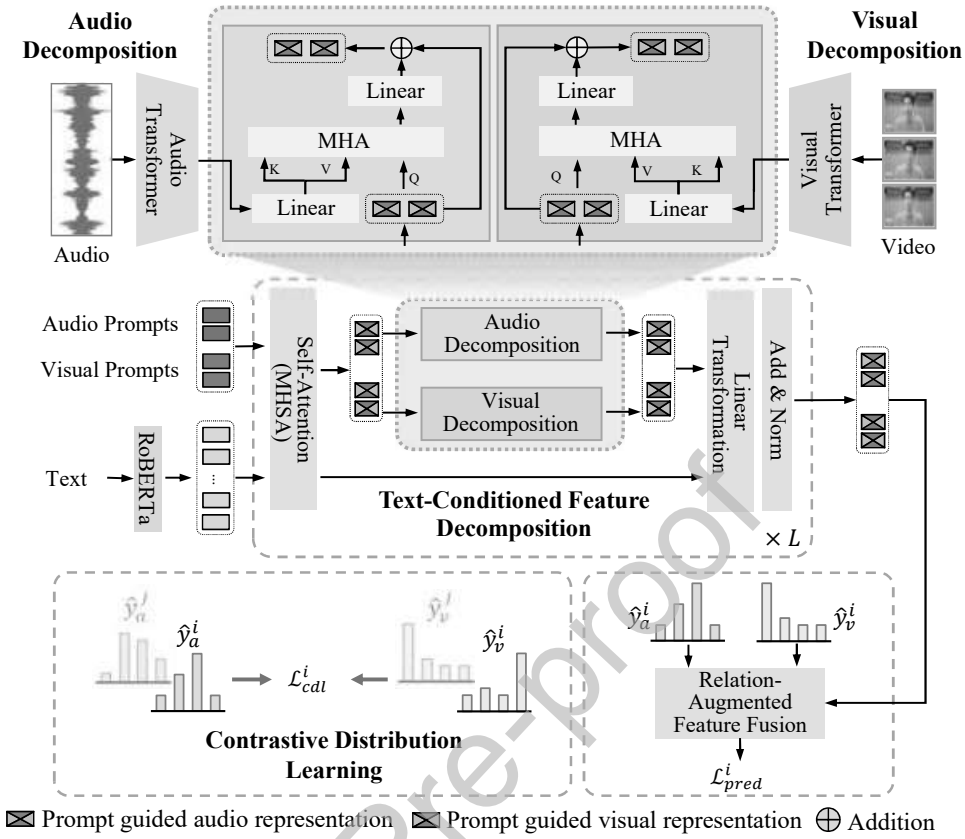

Figure 2: The structure of Prompt-based Distribution learning (PDL) framework. It has L text-conditioned feature decomposition layers. PDL consists of three components, including Text-Conditioned Feature Decomposition (TCFD), Contrastive Distribution Learning (CDL) and Relation-Augmented Feature Fusion (RAFF). "MHA" is the multi-head attention, $\mathcal{L}^i_{pred}$ means the Contrastive Distribution Learning loss and $\mathcal{L}^i_{cdl}$ means the prediction loss. $\hat{y}^i_h$ and $\hat{y}^j_h$ represent the sentiment predictions of modality h for sample i and sample j, respectively.

decision making process.

2. Related Work

2.1. Multimodal Sentiment Analysis

While significant attention has been given to fusion functions, many approaches overlook the inherent ambiguity in multimodal data. Several studies [14, 15] have emphasized that the text modality plays a pivotal role, often providing the most reliable semantic cues. We adopt this text-centric paradigm as the foundation for our framework. However, a critical challenge remains in how ambiguity is defined and addressed. Most prior works on uncertainty focus on aleatoric uncertainty, such as corrupted audio, overexposed images, or unreliable sensor readings [16]. In contrast, our work tackles a different and more complex challenge: semantic and annotation ambiguity. This includes explicit ambiguity from inter-modality conflicts, such as a sarcastic tone paired with positive words, and implicit ambiguity from the subjective nature of human annotation, where annotators' backgrounds can lead to varied labels for the same sample. To address these specific challenges, we propose Prompt-based Distribution Learning (PDL). Guided by the strong cues from the text modality, PDL introduces modality-aware prompts for the audio and visual modalities. While related work [17] ensembles Bayesian prompts to obtain a reliable weighted answer, our prompts act as probes to disentangle modality-specific semantics. Crucially, we then transform these discrete embeddings into a flexible sentiment distribution, simulating the varied judgments of human annotators.

2.2. Unimodal Representation Learning

Unimodal representation learning aims to extract high-quality, noise-free, and information-preserving features from a single modality, playing a crucial role in determining the final performance of sentiment recognition. In recent years, various approaches [3, 18–22] have been proposed to learn robust and sentiment-relevant unimodal representations. MISA [3] and MDSE [23] address this objective by decomposing unimodal features into a combination of modality-invariant and multiple modality-specific representations, enabling more disentangled and interpretable modeling of each modality. Self-MM [18] introduces a self-supervised learning framework that explores the semantic alignment between multimodal and unimodal data. It generates pseudo unimodal labels from multimodal signals, guiding the learning of discriminative unimodal representations under the constraint of semantic consistency. CLIP [19], on the other hand, constructs large-scale text-image pairs and learns to project different modalities into a shared joint embedding space. This approach significantly reduces the semantic gap between modalities and facilitates transferable feature learning. In this work, we argue that simply seeking shared semantics across modalities is insufficient, as these semantics are inherently complex and multimodal sentiment labels tend to be ambiguous. Therefore, we propose modality-aware prompts to guide unimodal representation learning, enabling a decoupling of modality-specific features under the guidance of the textual modality.

2.3. Label Distribution Learning

Label Distribution Learning (LDL) [24] provides a paradigm for addressing label ambiguity, describing an instance with a distribution over all possible labels rather than a single one. It has been widely applied in tasks with inherent ambiguity, such as age estimation [25] and facial expression recognition [26]. We leverage the core intuition of LDL as a theoretical foundation and explicitly model the subjective judgments and varying scores that different human annotators may assign. Our key innovation lies in how this distribution is generated. Many prior works rely on a pre-defined Gaussian distribution assumption to model the data noise[16] or the label ambiguity [26]. Our PDL framework, however, does not rely on this rigid assumption. We introduce a prompt-based decomposition mechanism that learns a flexible, data-driven sentiment distribution. By using prompts as probes to decouple features, we simulate the complex annotation process, rather than assuming a fixed distribution shape to sensor noise or label consensus.

3. Methodology

This section briefly introduces the characteristics of PDL, including its de-coupling and integration processes, as illustrated in Figure 2. Section 3.1 first briefly introduces the multimodal inputs of the task and the modality-aware prompts, followed by a detailed description of the PDL architecture and the sentiment distribution learning process for multimodal representations. Subsequently, Section 3.2 presents a detailed description of feature decomposition using modality-aware prompting. Section 3.3 describes the construction of the sentiment distribution. Then, Section 3.4 details the working mechanism of the relation-augmented feature fusion module. Finally, Section 3.5 introduces the model's optimization objectives, including the prediction loss and the contrastive distribution learning loss.

3.1. Multimodal Input and Modality-aware Prompt

For sample $\mathcal{X}^i = \{x_a^i, x_t^i, x_v^i, y^i\}$ consisting of three modalities, audio (a), textual (t) and visual (v), where $x_h^i \in R^{T_h \times D_h}, h \in \{a, t, v\}$ is the feature of h modality, y^i is the multimodal sentiment label of $\mathcal{X}^i$, T_h is the sequence length and D_h is the dimension. Our goal is to take x_a^i, x_t^i and x_v^i as inputs and output the predicted sentiment intensity $\hat{y}^i$. We first use transformers to encode audio and visual features, and then map all the multimodal features to a unified dimension D. Specifically, $x_a^i \in R^{T_a \times D}$, $x_t^i \in R^{T_t \times D}$ and $x_v^i \in R^{T_v \times D}$ represent the audio, textual and visual features, respectively. $T_h (h \in \{a, t, v\})$ denotes the sequence length of features, D is the sequence feature dimension.

To effectively compress audio and visual features and enable sufficient interaction with the textual modality, we initialize learnable modality-aware prompts for both the audio and visual modalities, respectively. Their mechanism is twofold: (1) They act as bridges that capture global sentiment context from text, using

it to guide the semantic decomposition of unimodal features and extract distinct sentiment cues. (2) The resulting decoupled representations are then used to construct the unimodal sentiment distributions, which are essential for the subsequent alignment and fusion stages. The audio prompt is denoted as $p_a \in R^{T_p \times D}$, and the visual prompt is represented as $p_v \in R^{T_p \times D}$. Here, T_p signifies the sequence length of the prompts.

Subsequently, we feed the textual, audio, and visual features, along with the modality-aware prompts, into the TCFD module to attend to relevant modality-specific cues and filter out irrelevant or noisy signals.

3.2. Feature Decomposition with Modality-aware Prompting

Modality-aware prompts are expected to bridge the semantic gap between different modalities, compress unimodal representations, and capture discrete sentiment semantics. However, the audio and visual modalities often contain noisy, ambiguous, or sentiment-irrelevant signals. In contrast, the textual modality typically provides the most reliable and explicit semantic cues. Therefore, to exploit these strong textual cues, we introduce the Text-Conditioned Feature Decomposition (TCFD) layers. Using a two-stage attention strategy, TCFD employs the modality-aware prompts as semantic probes. These probes guide unimodal disentanglement by alternately aligning semantics and selectively decomposing the audio and visual features into fine-grained, sentiment-relevant components. For the l-th layer, the prompts are first combined with textual features to activate textual sentiment information,

$$\begin{aligned} H^{i,l} &= [P_a^{i,l}, P_v^{i,l}, x_t^{i,l}], \\ [\hat{P}_a^{i,l}, \hat{P}_v^{i,l}, \hat{x}_t^{i,l}] &= \mathrm{MHSA}(H^{i,l}), \end{aligned} \tag{1}$$

where $P_a^{i,l}$ and $P_v^{i,l}$ are the prompts of the l-th layer and $x_t^{i,l}$ is the text features at l-th layer. $[\cdot,\cdot]$ is the concatenation. MHSA$(\cdot)$ stands for Multi-Head Self-Attention. $\hat{P}_a^{i,l}$ and $\hat{P}_v^{i,l}$ are the text-informed prompts. For the first layer, $P_a^{i,l} = p_a$, $P_v^{i,l} = p_v$, and $x_t^{i,l} = x_t^i$. Then, we utilize the text-informed prompts as query to decompose the audio and visual features separately.

Audio Decomposition. Given the audio features x_a^i, a linear transformation is first employed to transform it into the text feature space, yielding $\hat{x}_a^i$. Then, text-informed audio prompt $\hat{P}_a^{i,l}$ is taken as query ($\mathcal{Q}$), $\hat{x}_a^i$ is act as key ($\mathcal{K}$) and value ($\mathcal{V}$). The audio decomposition can be achieved by:

$$\begin{aligned} \hat{x}_a^i &= \mathrm{fc}(x_a^i), \\ A^{i,l} &= \mathrm{fc}(\mathrm{MHA}(\mathcal{Q}, \mathcal{K}, \mathcal{V})) + \mathcal{Q}, \end{aligned} \tag{2}$$

where $\mathrm{MHA}(\mathcal{Q}, \mathcal{K}, \mathcal{V})$ denotes the multi-head attention, which is designed to capture different patterns in audio features.

Visual Decomposition. Similar to the audio decomposition method, we use text-informative visual prompt $\hat{P}_v^{i,l}$ as query ($\mathcal{Q}$), and $\hat{x}_v^i$ as key ($\mathcal{K}$) and value ($\mathcal{V}$). The visual decomposition is formulated as:

$$\begin{aligned} \hat{x}_v^i &= \mathrm{fc}(x_v^i), \\ V^{i,l} &= \mathrm{fc}(\mathrm{MHA}(\mathcal{Q}, \mathcal{K}, \mathcal{V})) + \mathcal{Q}. \end{aligned} \tag{3}$$

We combine the decomposed multimodal features and apply a linear transformation with a residual connection to obtain final embeddings,

$$\begin{aligned} Z^{i,l} &= [A^{i,l}, V^{i,l}, \hat{x}_t^{i,l}] + H^{i,l}, \\ [P_a^{i,l+1}, P_v^{i,l+1}, x_t^{i,l+1}] &= \mathrm{LN}(\mathrm{fc}(Z^{i,l}) + Z^{i,l}), \end{aligned} \tag{4}$$

where LN$(\cdot)$ is layer normalization, $P_a^{i,l+1}$ and $P_v^{i,l+1}$ are the sentiment-oriented embeddings for audio and visual modalities. We perform L layers of TCFD to fully disentangle the audio and visual features. This allows $P_a^{i,L}$ and $P_v^{i,L}$ to aggregate sentiment-related semantics from different levels of each unimodal features.

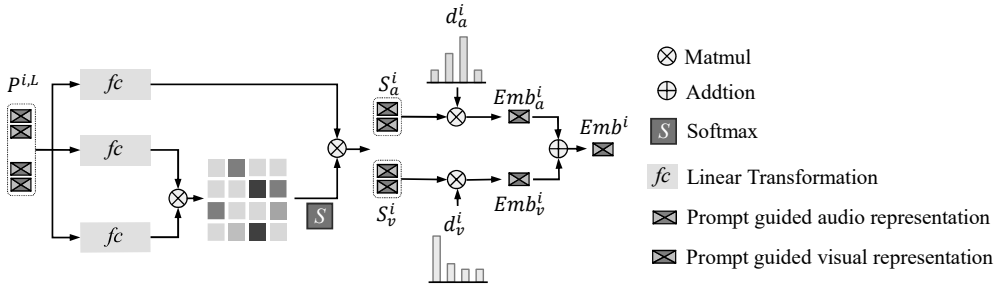

Figure 3: Architecture of the Relation-Augmented Feature Fusion (RAFF) module. RAFF first employs an attention mechanism to activate contextual information within the cross-modally decoupled embeddings. Then, it utilizes the semantic distribution as dynamic weights to fuse the unimodal high-level semantics with the integrated cross-modal semantics.

3.3. Sentiment Distribution

Following the operations above, $P_a^{i,L}$ and $P_v^{i,L}$ now contain the semantically disentangled information for the audio and visual modalities, respectively. Each embedding within them focuses on distinct sentimental traits in the features. We then simulate the annotation process of different annotators by deriving the sentiment intensity of each embedding through embedding-specific linear annotators, finally yielding the final sentiment distribution. The audio sentiment distribution can be achieved by:

$$\begin{aligned} Prob_a^i &= \text{Relu}(\text{MLP}(P_a^{i,L})), \\ d_a^i &= \text{Softmax}(Prob_a^i), \end{aligned} \tag{5}$$

and visual sentiment distribution can be achieved by:

$$\begin{aligned} Prob_v^i &= \text{Relu}(\text{MLP}(P_v^{i,L})), \\ d_v^i &= \text{Softmax}(Prob_v^i), \end{aligned} \tag{6}$$

where $d_a^i \in R^{T_p \times 1}$ and $d_v^i \in R^{T_p \times 1}$. In this way, the decoupled cross-modal features are converted into a semantic intensity distributions. For each sample, two distributions are derived through feature decomposition, representing the intensity of modality-specific semantics.

3.4. Relation-Augmented Feature Fusion

After the TCFD module, the decoupled embeddings, $P_a^{i,L}$ and $P_v^{i,L}$, represent fine-grained semantics. The structural relations across embeddings encode rich contextual information. Besides, the sentiment distributions, d_a^i and d_v^i quantify the semantic intensity included in each embedding. We utilize these relations to dynamically reinforce feature representations. As shown in Figure 3, given the audio embeddings $P_a^{i,L} \in R^{T_p \times D}$ and visual embeddings $P_v^{i,L} \in R^{T_p \times D}$, we employ a self-attention mechanism to exploit the multimodal relations,

$$\begin{aligned} E^i &= [P_a^{i,L}, P_v^{i,L}], \\ Att^{i,jk} &= \frac{f(e^{i,j}, e^{i,k})}{\sum_k f(e^{i,j}, e^{i,k})}, \\ f(e^{i,j}, e^{i,k}) &= \exp(fc(e^{i,j})^T \cdot fc(e^{i,k})). \end{aligned} \tag{7}$$

Then we adopt inner product of the calculated relations $Att^i \in R^{2T_p \times 2T_p}$ and embeddings $E^i \in R^{2T_p \times 1}$,

$$\hat{E}^i = \text{Matmul}(Att^i, fc(E^i)), \tag{8}$$

where $Att^{i,hk} \in Att^i$ is the calculated contextual relationship, $e^{i,j}, e^{i,k} \in E^i$. $\text{Matmul}(\cdot, \cdot)$ stands for matrix multiplication. We repeat above process by multiple times to fully utilize the multi-level contextual relationships. For convenience, we implement this module by Multi-Head Self-Attention mechanism, where $\hat{E}^i = \text{MHSA}(E^i)$. $\hat{E}^i$ are the enhanced representations.

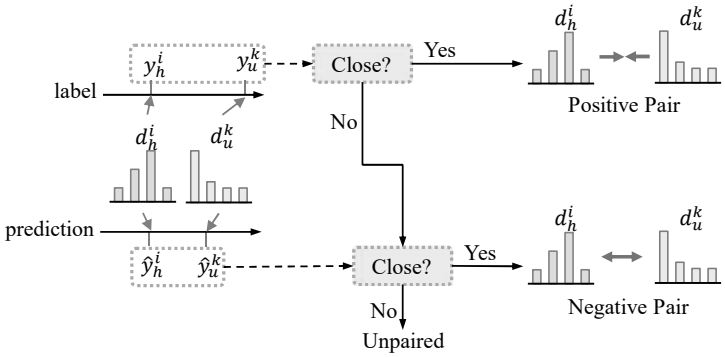

Figure 4: Overview of Contrastive Distribution Learning (CDL) loss. The CDL loss computes both cross-modal and unimodal semantic distribution losses in the same manner. For clarity, the figure illustrates the process of constructing positive and negative sample pairs based on distribution partitions from different modalities.

Sentiment distributions represent unimodal sentiment cues, reflecting both the semantic attributes and the intensity within the decoupled unimodal features. These distributions are used as weights to calculate a weighted sum of $\hat{E}^i$ as final representation,

$$\begin{aligned}[S_a^i, S_v^i] &= \hat{E}^i, \\ Emb^i &= \frac{1}{2}(d_a^i)^\top \cdot S_a^i + \frac{1}{2}(d_v^i)^\top \cdot S_v^i \\ &= \frac{1}{2}Emb_a^i + \frac{1}{2}Emb_v^i,\end{aligned} \tag{9}$$

where $S_a^i \in R^{T_p \times D}$ and $S_v^i \in R^{T_p \times D}$ are the enhanced audio and visual representations, respectively, Emb^i is the fused multimodal features. Then, linear regression layers are adopted to obtain the final predictions $\hat{y}^i = fc(Emb^i)$, $\hat{y}_h^i = fc(Emb_h^i)$, $h \in \{a, v\}$.

3.5. Optimization Objectives

For the sample $\mathcal{X}^i$, we employ prediction loss ($\mathcal{L}_{pred}^i$) and contrastive distribution learning loss ($\mathcal{L}_{cdl}^i$) to optimize the sentiment fusion and predictive ability of the model.

Prediction. We expect the unimodal features, extracted by the TCFD module and further refined by the RAFF module, to retain modality-specific semantics, while also aiming to accurately predict the overall multimodal sentiment intensity. To achieve this, we adopt the L1 paradigm for both unimodal and multimodal prediction. Specifically, we compute the unimodal loss and multimodal loss based on the predicted unimodal labels and multimodal labels, respectively. The sum of these losses are defined as the overall prediction loss:

$$\mathcal{L}_{pred}^i = \|\hat{y}^i - y^i\|_1 + \|\hat{y}_a^i - y_a^i\|_1 + \|\hat{y}_v^i - y_v^i\|_1, \tag{10}$$

here, y^i is the multimodal sentiment annotation, $\hat{y}^i$ is the multimodal sentiment prediction, y_h^i and $\hat{y}_h^i$ represent unimodal sentiment annotation and prediction, respectively, where $h \in \{a, v\}$. $\hat{y}^i$ denotes the final sentiment prediction result of the model.

Contrastive Distribution Learning. We aim for the learned sentiment distribution to accurately reflect the sentiment polarity of the features. The design of the CDL loss is based on a core intuition: samples with similar sentiment labels should also possess similar high-level sentiment distributions. Building on this intuition, we note that the local and global correlations between samples offer valuable insights for effectively modeling these relationships within the distribution space. Therefore, we propose a contrastive regularizer. This regularizer incorporates both unimodal and cross-modal contrastive learning losses to enforce this principle, preventing the distributions of irrelevant samples from collapsing into the same point and ensuring that semantically similar samples are pulled closer together. As illustrated in Figure 4, given

a distribution d_h^i with its unimodal label y_h^i, we construct a positive set $\mathcal{P}_h^i$ by sampling distributions from the distribution set $\mathcal{D} = \left\{d_u^k | i = 1, ..., B; u \in \{a, v\}\right\}$ in a bath whose true label y_u^k difference from y_h^i is less than a threshold σ, where B is batch size. Conversely, if the absolute difference in true labels is no less than σ, but the predicted labels are undesirably similar (i.e., closer than σ), the corresponding distribution is included in the negative set $\mathcal{N}_h^i$,

$$\mathcal{P}_h^i = \left\{d_u^k \in \mathcal{D} \middle| \ |y_h^i - y_u^k| < \sigma, \ i \neq k \ or \ h \neq u\right\}, \tag{11}$$

$$\mathcal{N}_h^i = \left\{d_u^k \in \mathcal{D} \middle| \ |y_h^i - y_u^k| \geq \sigma, |\hat{y}_h^i - \hat{y}_u^k| < \sigma, \ i \neq k \ or \ h \neq u\right\}, \tag{12}$$

where $h, u \in \{a, v\}$. Thus, $\mathcal{P}_h^i$ and $\mathcal{N}_h^i$ contain distribution from both the same modality as d_h^i and other modality. The contrastive distribution learning loss is:

$$\begin{aligned}\mathcal{L}_{cdl}^i = -log\frac{1}{|\mathcal{P}_a^i|}\sum_{d_u^j \in \mathcal{P}_a^i}\frac{f(d_a^i, d_u^j)}{\sum\limits_{d_h^g \in \mathcal{P}_a^i} f(d_a^i, d_h^g) + \sum\limits_{d_q^k \in \mathcal{N}_a^i} P_{a,q}^{i,k} f(d_a^i, d_q^k)} \\ - log\frac{1}{|\mathcal{P}_v^i|}\sum_{d_u^j \in \mathcal{P}_v^i}\frac{f(d_v^i, d_u^j)}{\sum\limits_{d_h^g \in \mathcal{P}_v^i} f(d_v^i, d_h^g) + \sum\limits_{d_q^k \in \mathcal{N}_v^i} P_{v,q}^{i,k} f(d_v^i, d_q^k)},\end{aligned} \tag{13}$$

where $|\mathcal{P}_h^i|$ is the size of set $\mathcal{P}_h^i$. $f(d_h^i, d_u^j)$ is the similarity between distribution d_h^i and d_u^j, $P_{h,u}^{i,k} = \exp(|y_h^i - y_u^k|)$ is used to dynamically punish negative pairs, where $h, u \in \{a, v\}$. The objective of CDL loss is to enforce similarity between semantically consistent distributions, i.e. maximizing the similarity of distribution pairs in $\mathcal{P}_a^i$ and $\mathcal{P}_v^i$, while penalizing the similarity of distributions associated with incorrect predictions, i.e. minimizing the similarity of distribution pairs in $\mathcal{N}_a^i$ and $\mathcal{N}_v^i$. In the experiment, we use the Earth Mover's Distance (EMD) [27] to measure the distance between distributions,

$$f(d_h^i, d_u^j) = \exp(-\text{EMD}(d_h^i, d_u^j)). \tag{14}$$

Finally, we refine the sentiment distributions and multimodal embeddings by optimizing a multi-task learning objective function:

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}(\mathcal{L}_{pred}^i + \lambda \cdot \mathcal{L}_{cdl}^i), \tag{15}$$

where B is the batch size, λ is a the loss weights to balance the gradients of the two losses.

4. Experiments

A comprehensive experimental investigation is conducted to evaluate the effectiveness of the proposed PDL model. This section details the experimental setup, encompassing datasets in Section 4.1, raw feature extraction in Section 4.2, model training details in Section 4.3, baseline methods in Section 4.4 and the evaluation metrics of experiments in Section 4.5. In Section 4.6, a comparative analysis is conducted with other competing methods. Section 4.7 provides a series of ablation studies to examine the role of key components and different modalities. Section 4.8 explores the influence of several key hyperparameters in PDL. Section 4.9 provides an analysis of the model's computation efficiency. Section 4.10 provides qualitative insights into the feature distributions and the learned sentiment distributions. We also randomly select three samples to demonstrate the effectiveness of the PDL model in Section 4.12. Finally, Section 4.13 demonstrates the generalizability of the PDL by conducting cross-dataset and cross-lingual evaluations.

Dataset	Train	Valid	Test	Total	Language
CMU-MOSI	1284	229	686	2199	English
CMU-MOSEI	16326	1871	4659	22856	English
CH-SIMS	1368	456	457	2281	Chinese

Table 1: The statistics of MOSI, MOSEI and CH-SIMS.

Dataset	CMU-MOSI	CMU-MOSEI	CH-SIMS
Batch Size B	32	64	64
Learning Rate	1e-4	1e-4	1e-4
Optimizer	Adam	Adam	Adam
Epochs	80	80	80
Depth L	2	2	2
T_p	8	6	8
Dimension D	512	256	256
λ	0.1	0.1	0.1
σ	0.1	0.1	0.1

Table 2: Hyperparameter settings on CMU-MOSI, CMU-MOSEI and CH-SIMS.

4.1. Datasets

We evaluate PDL on three public datasets, CMU-MOSI, CMU-MOSEI and CH-SIMS. The statistics of datasets can be seen in Table 1.

CMU-MOSI [12] is a CMU multimodal opinion-level sentiment intensity dataset. It is an English dataset derived from YouTube, the length of the videos range from 2 to 5 minutes. These videos are split into 2199 video clips, each clip is annotated with a continuous multimodal sentiment score from -3 (strongly negative) to 3 (strongly positive).

CMU-MOSEI [13] is a CMU English multimodal opinion sentiment and sentimental intensity dataset. It is composed of 23453 video utterances with 250 topics from 1000 distinct speakers. Each utterance is annotated with a multimodal sentiment score from -3 (strongly negative) to 3 (strongly positive).

CH-SIMS [4] is a Chinese multimodal sentiment analysis dataset. It contains 60 videos from movies, TV series and variety shows. These videos are split into 2280 segments, with an average segment length of 3.67 seconds. Each segment is annotated not only with a continuous multimodal sentiment score, but also with unimodal sentiment scores for each modality, ranging from -1 (strongly negative) to 1 (strongly positive).

4.2. Raw Feature Extraction

The raw feature extraction follows common practice in Multimodal Sentiment Analysis [5, 14, 28–30]. For the text modality, the pre-trained RoBERTa [31] is applied to extract text features. For the audio modality, the librosa [32] is used to extract acoustic features, including 12 Mel Frequency Cepstral Coefficients (MFCC), Zero-Crossing Rate (ZCR) and Chromagram. For the visual modality, the open-source OpenFace2 [33] toolkit is employed to extract facial features, including eye movement, facial keypoints, and facial expressions.

4.3. Model Training Details

During training, we use the Adam[34] optimizer and train models on a single NVIDIA RTX 4090 GPU. The model is trained for 80 epochs to ensure adequate convergence. The hyper-parameters of the model are determined based on ablation studies results, with the weighting factor $\lambda = 0.1$ across all three datasets. For the dataset CMU-MOSI, we set the the batch size to 32, the sequence length of prompt $T_p = 8$, $D = 512$, $L = 2$ and $\sigma = 0.1$. For CMU-MOSEI, we set batch size to 64, $T_p = 6$, $D = 256$, $L = 2$ and $\sigma = 0.1$.

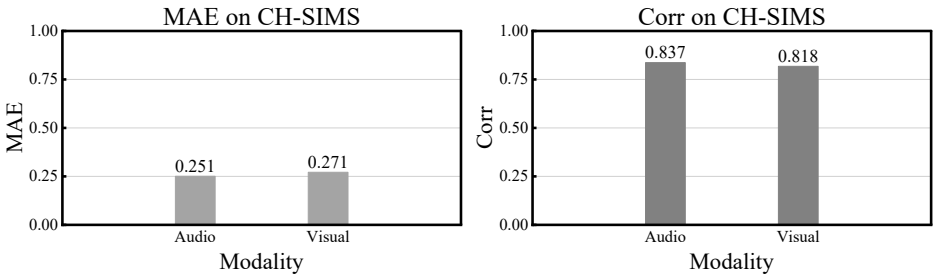

Figure 5: Performance of ULGM in unimodal label generation on the CH-SIMS dataset.

For CH-SIMS, we set batch size to 64, $T_p = 8$, $D = 512$, $L = 2$ and $\sigma = 0.1$. More details are provided in Table 2. To investigate the impact of batch size B and dimension D on the results, we also conducted additional experiments under two settings, ($B = 32$, $D = 512$) and ($B = 64$, $D = 256$).

Our framework utilizes unimodal labels for optimization. The CH-SIMS dataset provides ground-truth unimodal annotations, whereas the CMU-MOSI and CMU-MOSEI datasets lack them. Prior works typically address this challenge via two strategies: substituting the multimodal label for the unimodal ones [5, 28], or generating pseudo-unimodal labels. A common method for the latter is the ULGM module, a parameter-free and widely adopted model that generates unimodal annotations based on multimodal labels and modality features [18, 35]. We evaluate the performance of ULGM on CH-SIMS, utilizing its available unimodal ground-truth labels. As shown in Figure 5, ULGM demonstrates effective generation performance, achieving an MAE of 0.251 and a correlation of 0.837 for the audio modality. To provide a comprehensive evaluation, we conduct experiments under both configurations:

- **PDL^M**: This setting utilizes multimodal labels as direct substitutes for unimodal labels when computing the unimodal prediction loss and the CDL loss, i.e., $y_a^i = y^i$ and $y_v^i = y^i$.
- **PDL^U**: This setting leverages unimodal labels for both the prediction loss and the CDL loss. For the CH-SIMS dataset, we use its provided ground-truth unimodal labels. For CMU-MOSI and CMU-MOSEI, we employ the ULGM module to generate the pseudo-unimodal labels, formulated as $y_h^i = \text{ULGM}(y^i, Emb^i, Emb_h^i)$, where $h \in \{a, v\}$.

4.4. Baselines

To evaluate the relative performance of PDL, we benchmark our model against several prominent baselines, including:

MISA[3]: assumes that multimodal data can be decomposed and mapped into a modality-invariant subspace and multiple modality-specific subspaces. It introduces a multi-task loss to constrain the encoders, enabling them to capture a holistic view of multimodal features.

Self-MM[18]: augments the unimodal data annotations using a self-supervised approach and designs multi-modal and uni-modal tasks to learn the relationships between multi-modal and uni-modal representations.

SSLMM[36]: constructs a multimodal graph encoder to address the challenges of modality and label missingness in multimodal data. In the self-supervised pre-training stage, it employs contrastive learning to encourage similarity between two augmented graphs from the same sample while distinguishing graphs from different samples. In the fine-tuning stage, it mitigates the model's reliance on labeled data through semi-supervised learning and modality reconstruction.

DEVA[37]: introduces an Emotional Description Generator that transforms raw audio and visual data into textual sentiment descriptions, facilitating the capture of fine-grained sentiment embedded in visual and auditory modalities. In addition, a progressive fusion module is designed to gradually integrate visual and audio features for enhanced multimodal representation learning.

ConFEDE[5]: decomposes different unimodal data into similarity features and dissimilarity features, and then establishes contrastive learning based on the relationships between these features to model the connections between different modality data.

ALMT[14]: introduces an additional hypermodality representation to aggregate multimodal features and uses the text modality as a guide and designs a Transformer-based fusion structure to learn an irrelevance and conflict-suppressing representation from visual and audio modalities.

CRNet[38]: designs a text-centered cross-modal interaction network and uses gradient-based representation enhancement module to exploit the relationships of modality-invariant and modality-specific representation spaces.

MCL-MCF[39]: introduces a progressive fusion framework that combines multi-level contrastive learning and multi-layer convolution fusion to gradually align and integrate multimodal features, effectively reducing modality heterogeneity.

MFON[40]: designs a text-guided cross-modal information interaction module and uses prompts to balance the optimization of different modalities, effectively extracting meaningful information from each modality.

DFMU[7]: is designed to address subjective bias in the annotation process and the complex set relationships of sentiment features using Gaussian distributions. The method models uncertainty by encoding a probabilistic distribution and adaptively adjusting the optimization objective. It also introduces a distribution-based contrastive learning mechanism to capture the uncertainty relationships between features.

ULMD[41]: decomposes modality representations into modality-invariant and modality-specific components. It then combines the modality-invariant features of text with those of vision and audio, respectively, to interpret sentiment semantics in terms of sentiment polarity and sentiment intensity. A multi-task learning framework is subsequently constructed by incorporating the ULGM module to generate unimodal labels.

MMoLRE[29]: proposes a unimodal task-specific feature extraction module that disentangles task-specific features and models inter-task correlations. Specifically, it introduces shared and task-specific experts to jointly model common and unique tasks across multiple learning objectives, thereby mitigating conflicts caused by task correlation.

Semi-IIN[30]: utilizes semi-supervised learning to reduce its dependence on expensive multimodal sentiment data. Simultaneously, it employs masked attention and gating mechanisms to dynamically select more important information and reduce interference.

4.5. Evaluation Metrics

Following the approaches of Zhang et al. [14] and Yang et al. [5], we use both regression and multi-class classification metrics. For regression, we report the Mean Absolute Error (MAE) and Pearson Correlation Coefficient (Corr). For classification, we report the multi-class accuracy and weighted F1-score. For CMU-MOSI and CMU-MOSEI datasets, we calculate the accuracy of 2-class prediction (Acc-2) and 7-class prediction (Acc-7). It should be noted that Acc-2 accuracy and F1-score can be computed in two ways: negative/non-negative (including zero) and negative/positive (exclude zero). For the CH-SIMS dataset, we report classification accuracy for 2-class prediction (Acc-2), 3-class prediction (Acc-3) and 5-class prediction (Acc-5). Except for MAE (lower is better), higher values indicate better performance for all other metrics.

4.6. Results

The performances comparison between all baselines and PDL on CMU-MOSI, CMU-MOSEI and CH-SIMS are summarized in Table 3 and Table 4. The reported results are averaged over 5 runs.

As shown in Table 3, given the absence of unimodal labels in the CMU-MOSI and CMU-MOSEI datasets, we report experimental results for two versions of our model: (1) PDL^M, which utilizes multimodal labels as substitutes for unimodal labels when computing the unimodal prediction loss and the CDL loss; and (2) PDL^U, which employs the ULGM module to generate pseudo-unimodal labels, subsequently using these generated labels for the unimodal prediction loss and the CDL loss.

Model	CMU-MOSI					CMU-MOSEI				
	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
MISA(2020)	81.80/83.40	81.70/83.60	42.30	0.783	0.761	83.60/85.50	83.80/85.30	52.20	0.555	0.756
Self-MM(2021)	83.44/85.46	83.36/85.43	46.67	0.708	0.796	83.76/85.15	83.82/84.90	53.87	0.530	0.764
SSLMM(2025)	83.40/85.10	83.50/85.40	46.80	0.703	0.798	83.60/86.00	83.70/86.10	53.30	0.530	0.773
DEVA(2025)	84.40/86.29	84.48/86.30	46.32	0.730	0.787	83.26/86.13	82.93/86.21	52.26	0.541	0.769
ALMT(2023)	84.55/86.43	84.57/86.47	49.42	0.683	0.805	84.78/86.79	85.19/86.86	54.28	0.526	0.779
ConFEDE(2023)	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	54.86	0.522	0.780
ULMD(2025)	-/85.82	-/85.71	47.81	0.700	0.799	-/85.95	-/85.91	53.81	0.531	0.770
CRNet(2024)	-/86.40	-/86.40	47.40	0.712	0.797	-/86.20	-/86.10	53.80	0.541	0.771
MCL-MCF(2024)	84.90/87.30	84.70/87.20	-	0.692	0.799	84.20/86.40	84.40/86.30	-	0.536	0.767
MFON(2025)	84.84/86.89	84.75/86.86	44.90	0.725	0.797	82.70/86.32	83.13/86.29	53.72	0.528	0.780
DFMU(2025)	85.78/87.66	86.01/87.68	49.45	0.663	0.830	85.65/87.59	85.61/87.62	54.50	0.513	0.780
MMoLRE(2025)	85.96/88.21	85.93/**88.23**	47.52	0.666	**0.837**	86.31/86.57	86.19/86.28	55.78	0.505	0.797
Semi-IIN(2025)	85.28/87.04	85.19/87.00	46.50	0.679	0.822	84.98/87.70	85.27/87.65	55.89	**0.497**	**0.804**
PDL^M($B=32$, $D=512$)	**86.15**/87.96	**86.11**/87.96	50.00	**0.636**	0.832	86.31/87.34	86.29/87.12	53.19	0.511	0.797
PDL^M($B=64$, $D=256$)	85.71/87.65	85.64/87.62	**50.87**	**0.636**	**0.837**	86.16/87.31	86.29/87.22	54.35	0.517	0.789
PDL^U($B=32$, $D=512$)	85.57/**88.26**	85.42/88.20	49.27	0.641	0.825	85.96/**87.92**	86.08/**87.78**	55.76	0.500	0.801
PDL^U($B=64$, $D=256$)	85.42/87.36	85.33/87.31	49.13	0.644	0.827	**86.56**/87.34	**86.59**/87.18	**56.19**	**0.497**	0.801

Table 3: Results on CMU-MOSI and CMU-MOSEI. In Acc-2 and F1 score, the left of the '/' corresponds to "negative/non-negative" and right corresponds to "negative/positive". '↑' indicates better performance as its value increases, '↓' indicates better performance as its value decreases. "PDL^M" utilizes multimodal labels as substitutes for unimodal labels when computing the unimodal prediction loss and the CDL loss. "PDL^U" employs the ULGM module to generate pseudo-unimodal labels, subsequently using these generated labels for the unimodal prediction loss and CDL loss. Best results are in bold; results surpassing baselines are underlined.

On the CMU-MOSI dataset, our method achieves competitive performance across multiple metrics, with notable results in Acc-7 and MAE. Specifically, the PDL^M variant attains a competitive Acc-7 of 50.00%, surpassing the previous best ALMT by 0.58%. Moreover, both PDL^M and PDL^U also perform strongly in MAE, with PDL^M achieving 0.636, a reduction of 0.027 compared to the strong baseline DFMU. Similarly, experiments on the CMU-MOSEI dataset show that PDL^U achieves favorable results on fine-grained metrics, including Acc-7, MAE, and Corr. This evidence collectively illustrates the advantages of our approach over the compared methods. We suggest that our proposed use of modality-aware prompts to decouple unimodal semantics enables a more precise mining of latent sentiment distributions within each modality, thereby enhancing the model's overall representational power.

Furthermore, we conduct experiments on the CH-SIMS dataset, which provides ground-truth unimodal annotations. As detailed in Table 4, we define two experimental setups: (1) PDL^M, using multimodal labels in place of unimodal ones; and (2) PDL^U, which leverages the ground-truth unimodal labels for the unimodal prediction and CDL losses.

The results indicate that our method achieves highly competitive performance under both configurations. In the PDL^M setup, the model achieves a SOTA Acc-5 of 47.70%, exceeding the next-best baseline, ConFEDE (46.30%). In the PDL^U setup, where ground-truth labels are used, our model attains SOTA results across four key metrics: Acc-2, F1-score, MAE, and Corr. Notably, the Corr score reaches 0.655, a significant improvement of 0.028 over DFMU, another method utilizing Gaussian distributions to address subjective annotation bias. These findings robustly demonstrate our method's ability to achieve outstanding results across diverse datasets, validating its effectiveness.

We found that the optimal hyperparameters, such as batch size B and dimension D, vary across datasets. This is expected, as the benchmarks have distinct properties, e.g., data source, sample size, language, and optimal settings were determined empirically for each. Crucially, our analysis also shows that a given experimental setup, i.e., PDL^M or PDL^U, variations in these parameters, such as ($B=32$, $D=512$) and ($B=64$, $D=256$) result in only minor performance differences. This suggests the slight fluctuations are

Model	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
MFON(2025)	78.56	78.51	-	-	0.420	0.594
DEVA(2025)	79.64	80.32	65.42	43.07	0.424	0.583
Self-MM(2021)	80.04	80.44	65.47	41.53	0.425	0.595
SSLMM(2025)	80.06	80.56	65.48	39.52	0.409	0.589
CRNet(2024)	80.70	80.70	-	-	0.416	0.628
ALMT(2023)	81.19	81.57	68.93	45.73	0.404	0.619
DFMU(2025)	81.32	81.91	68.96	46.13	0.399	0.627
ConFEDE(2023)	82.23	82.08	**70.15**	46.30	0.392	0.637
PDL^M($B=32$, $D=512$)	80.09	80.15	67.83	**47.70**	**0.385**	0.637
PDL^M($B=64$, $D=256$)	80.31	80.52	68.71	47.05	0.391	0.631
PDL^U($B=32$, $D=512$)	**82.49**	82.56	69.80	47.05	0.400	0.616
PDL^U($B=64$, $D=256$)	**82.49**	**82.71**	69.15	46.17	0.390	**0.655**

Table 4: Results on CH-SIMS. '↑' indicates better performance as its value increases; '↓' indicates better performance as its value decreases. "PDL^M" utilizes multimodal labels as direct substitutes for unimdoal labels when computing the unimodal prediction loss and CDL loss. "PDL^U" leverages the ground-truth unimodal labels for the unimodal prediction and CDL losses. Best results are in bold; results surpassing baselines are underlined.

Components			CMU-MOSI					CMU-MOSEI				
TCFD	RAFF	CDL	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
			83.24/85.67	82.81/85.36	44.46	0.714	0.793	84.33/86.16	84.53/86.04	52.07	0.536	0.790
✓			83.97/85.82	83.92/85.83	46.36	0.683	0.814	85.98/87.04	86.04/86.87	53.36	0.523	0.794
✓	✓		84.11/86.28	84.04/86.27	46.65	0.718	0.785	86.39/87.23	86.35/86.99	53.12	0.513	0.795
✓	✓	✓	**85.57/88.26**	**85.42/88.20**	**49.27**	**0.641**	**0.825**	**86.56/87.34**	**86.59/87.18**	**56.19**	**0.497**	**0.801**

Table 5: Ablation results of the PDL^U model components on CMU-MOSI and CMU-MOSEI. (1) TCFD + RAFF + CDL: The complete PDL model; (2) TCFD + RAFF: PDL without the CDL loss, keeping only the multimodal prediction loss; (3) TCFD: Further removing the RAFF module, where unimodal representations are fused via simple average pooling, while still optimizing the multimodal prediction loss; 4) Plain: Direct concatenation the "CLS" tokens of unimodal representations without any fusion mechanism, using only the multimodal prediction loss for training.

Components			CH-SIMS					
TCFD	RAFF	CDL	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
			78.56	79.04	68.49	**48.36**	**0.390**	0.614
✓			80.31	80.52	67.61	45.30	0.397	0.633
✓	✓		80.53	80.62	66.52	45.08	0.392	0.632
✓	✓	✓	**82.49**	**82.71**	**69.15**	46.17	**0.390**	**0.655**

Table 6: Ablation results of the PDL^U model components on CH-SIMS.

likely due to computational stochasticity, confirming that our model is robust and not overly sensitive to these specific hyperparameter choices.

A comparative analysis of the PDL^M and PDL^U results on CH-SIMS reveals that PDL^U with true labels outperforms PDL^M on multiple metrics. This suggests a potential inconsistency between multimodal and unimodal annotations, which may introduce semantic conflict and noise. It highlights that accurate unimodal labels can more effectively guide feature learning and semantic mining. Nevertheless, it is crucial to note that our model achieves SOTA results across all datasets, including CH-SIMS under both setups and CMU-MOSI and CMU-MOSEI using generated or substituted labels. This strongly suggests that our model's architecture has a high degree of robustness to label noise.

4.7. Ablation Study and Analysis

We perform ablation studies to analyze how various configurations influence the overall performance of the PDL^U model. Specifically, we examine the impact of the number of TCFD layers, the modality-aware prompts, the RAFF module, the CDL loss function, and the contribution of each modality.

As shown in Table 5 and Table 6, removing any component led to a decrease in performance compared to the full model. Notably, the CDL loss significantly enhances fine-grained metrics such as MAE and Corr, demonstrating its effectiveness in better constraining semantic distributions and capturing subtle sentiment representations. Furthermore, each proposed component contributes to the model's capacity to understand complex cross-modal sentiment to varying degrees, highlighting their essential roles in improving overall performance. We further investigate the role of each component in detail.

Effects of TCFD Layer. The TCFD layer serves as the core component of PDL. When employing the TCFD layer independently, we concatenate all extracted embeddings and utilize a linear regression layer to predict sentiment polarity. As demonstrated in Table 5 and Table 6, the incorporation of the TCFD layer significantly enhances classification accuracy. The extraction of multiple features enables the model to more precisely capture of the sentiment information embedded within multimodal signals. Similarly, performance improvements are also observed in the regression task. To further assess the effectiveness of the TCFD layer, we conduct ablation studies on the final model. Results are shown in Table 7 and Table 8. We evaluate the effects of audio decomposition ("AD"), visual decomposition ("VD") and Multi-Head Self-Attention ("MHSA") individually. Most metrics degrade significantly when either "AD" or "VD" is removed. This indicates that both audio and visual modalities are important, as either modality can suffer from explicit ambiguity. Furthermore, the removal of MHSA leads to a significant degradation in model performance. These results indicate that MHSA provides crucial textual guidance, which is essential for effectively activating and decoupling unimodal sentiment semantics in subsequent stages.

Effects of Modality-aware Prompts. To gain deeper insights, we further examine the structure and role of the modality-aware prompts. We compare the modality-shared prompts ("Shared") with modality-aware prompts ("Private"), as shown in Table 9 and Table 10. The use of shared prompts leads to a notable performance drop, indicating the existence of explicit ambiguity between modalities. Shared prompts fail to effectively bridge the modality gap, whereas modality-aware prompts better adapt to the unique characteristics of each modality and extract higher-quality unimodal features under the guidance of the text modality.

Effects of RAFF module. As a feature fusion module, RAFF is designed to preserve core information while potentially disregarding some finer details. As illustrated in Table 5 and Table 6, RAFF demonstrates its effectiveness more on binary classification tasks compared to regression tasks. Regression tasks focus on predicting a precise value, thus necessitating the capture of fine-grained information. Conversely, classification tasks aim to achieve results within a predefined margin of error, which are generally easier to model.

Effects of CDL Loss. As shown in Table 5 and Table 6, the Contrastive Distribution Learning (CDL) loss significantly improves performance, especially in terms of Corr. CDL leverages sentiment distributions from audio and visual modalities, thereby exerting a substantial effect on extracting fine-grained details. More importantly, distribution learning effectively mitigates both explicit and implicit ambiguities by modeling intra-modal and inter-modal correlations. As shown in Table 11 and Table 12, we evaluate the impact of unimodal contrastive learning ($\mathcal{L}_{uc}$) and cross-modal contrastive learning ($\mathcal{L}_{cc}$). Specifically, Table 11 shows that removing $\mathcal{L}_{uc}$ causes a significant drop in performance. This is because unimodal contrastive learning facilitates cross-sample learning, enhancing multimodal representations by penalizing incorrect similarities.

Effects of Each Modality. In Table 13 and Table 14, we examine how modality omissions affect the performance of the PDL framework. Specifically, we examine two scenarios: text+audio (T+A) and text+visual (T+V). The results reveal that both audio and visual modalities contribute significantly to model performance, with the effect being particularly pronounced on the CH-SIMS dataset. Notably, however, on the CMU-MOSI dataset, the PDL model without the visual modality outperforms its multimodal counterpart in terms of MAE and Corr metrics. Similarly, on the CMU-MOSEI dataset, the "without neutral" setting achieves the highest results for the Acc-2 and F1-score metrics. These results suggest that across these

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
w/o MHSA AD	44.75/42.23	27.67/25.07	15.45	1.433	0.121	71.02/62.85	58.99/48.51	41.36	0.838	0.053
w/o MHSA VD	44.75/42.23	27.67/25.07	15.45	1.426	0.103	59.18/61.72	60.98/62.17	41.19	0.810	0.237
w/o MHSA	55.25/57.77	39.32/42.31	15.45	1.410	0.044	68.49/65.02	65.48/60.54	41.32	0.812	0.234
w/o AD	82.80/85.21	82.64/85.14	45.34	0.695	0.800	86.33/86.68	86.19/86.37	55.42	0.499	**0.806**
w/o VD	83.24/85.37	83.13/85.33	47.23	0.667	**0.828**	84.76/**87.67**	85.09/**87.64**	54.63	0.504	0.799
PDL^U	**85.57**/**88.26**	**85.42**/**88.20**	**49.27**	**0.641**	0.825	**86.56**/87.34	**86.59**/87.18	**56.19**	**0.497**	0.801

Table 7: The ablation results of TCFD layer on CMU-MOSI and CMU-MOSEI. "AD": the audio decomposition; "VD": the visual decomposition; "MHSA": the Multi-Head Self-Attention mechanism.

Method	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
w/o MHSA AD	69.37	56.82	54.27	21.23	0.588	0.180
w/o MHSA VD	69.37	56.82	54.27	21.23	0.586	0.164
w/o MHSA	69.37	56.82	54.27	21.23	0.588	0.039
w/o AD	77.68	78.03	66.96	42.89	0.416	0.591
w/o VD	77.90	78.05	66.52	43.76	0.407	0.607
PDL^U	**82.49**	**82.71**	**69.15**	**46.17**	**0.390**	**0.655**

Table 8: The ablation results of TCFD layer on CH-SIMS. "AD": the audio decomposition; "VD": the visual decomposition; "MHSA": the Multi-Head Self-Attention mechanism.

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
Shared	84.55/86.74	84.44/86.69	**49.85**	0.654	0.814	83.34/87.12	83.78/87.12	54.84	0.503	0.796
Private	**85.57**/**88.26**	**85.42**/**88.20**	49.27	**0.641**	**0.825**	**86.56**/**87.34**	**86.59**/**87.18**	**56.19**	**0.497**	**0.801**

Table 9: The ablation results of modality-aware prompts on CMU-MOSI and CMU-MOSEI. The experiments are based on the PDL^U model. "Shared" denotes the scenario where both audio and visual modalities share the same prompts, whereas "Private" represents our approach using modality-aware prompts.

Method	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
Shared	77.68	78.15	66.52	45.30	0.398	0.622
Private	**82.49**	**82.71**	**69.15**	**46.17**	**0.390**	**0.655**

Table 10: The ablation results of modality-aware prompts on CH-SIMS. The experiments are based on the PDL^U model. "Shared" denotes the scenario where both audio and visual modalities share the same prompts, whereas "Private" represents our approach using modality-aware prompts.

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
$\mathcal{L}_{uc}+\mathcal{L}_{cc}$	**85.57**/**88.26**	**85.42**/**88.20**	49.27	**0.641**	0.825	**86.56**/**87.34**	**86.59**/**87.18**	**56.19**	**0.497**	**0.801**
w/o $\mathcal{L}_{uc}$	83.82/85.82	83.68/85.74	49.42	0.666	0.816	85.36/86.41	85.43/86.23	55.44	0.510	0.793
w/o $\mathcal{L}_{cc}$	84.99/87.20	84.94/87.21	**49.56**	0.645	**0.826**	85.58/87.01	85.69/86.87	55.51	0.500	0.797

Table 11: The ablation results of unimodal and cross-modal CDL loss on CMU-MOSI and CMU-MOSEI. The experiments are based on the PDL^U model. $\mathcal{L}_{uc}$ means unimodal CDL loss, while $\mathcal{L}_{cc}$ means cross modality CDL loss.

Method	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
$\mathcal{L}_{uc}+\mathcal{L}_{cc}$	**82.49**	**82.71**	**69.15**	46.17	**0.390**	**0.655**
w/o $\mathcal{L}_{uc}$	**82.49**	82.56	68.05	**48.14**	0.392	0.654
w/o $\mathcal{L}_{cc}$	81.62	81.72	67.40	46.61	**0.390**	0.653

Table 12: The ablation results of unimodal and cross-modal CDL loss on CH-SIMS. The experiments are based on the PDL^U model. $\mathcal{L}_{uc}$ means unimodal CDL loss, while $\mathcal{L}_{cc}$ means cross modality CDL loss.

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
T+A	84.84/87.35	84.67/87.26	46.36	**0.637**	**0.836**	84.31/86.79	84.66/86.78	55.63	0.500	0.799
T+V	82.51/84.91	82.34/84.83	46.50	0.739	0.759	85.92/**88.00**	86.16/**87.95**	55.42	0.499	**0.802**
T+A+V	**85.57**/**88.26**	**85.42**/**88.20**	**49.27**	0.641	0.825	**86.56**/87.34	**86.59**/87.18	**56.19**	**0.497**	0.801

Table 13: The results of removing audio or video modalities on CMU-MOSI and CMU-MOSEI. The experiments are based on the PDL^U model, which utilizes unimodal labels generated by ULGM to compute the prediction loss and the CDL loss.

Method	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
T+A	78.99	79.32	65.86	44.20	0.399	0.612
T+V	79.43	79.81	65.21	42.01	0.402	0.616
T+A+V	**82.49**	**82.71**	**69.15**	**46.17**	**0.390**	**0.655**

Table 14: The results of removing audio or video modalities on CH-SIMS. The experiments are based on the PDL^U model, which utilizes ground-truth unimodal labels to compute the prediction loss and the CDL loss.

datasets, the text modality serves as the primary determinant of sentiment analysis outcomes, aligning with prior studies[14, 28]. Further analysis indicates that label uncertainty introduced by the ULGM may impede the effectiveness of PDL.

Comparison of Different Sequence Aggregation Strategies. We examine different sequence aggregation strategies for the decoupled features $P_a^{i,L}$ and $P_v^{i,L}$, including Global Average Pooling (GAP), Max Pooling (GMP) and attention-based methods that perform weighted aggregation using scores derived from linear annotators of varying granularity: MS-Attn, which uses a globally shared linear annotator, and MP-Attn, which relies on modality-specific linear annotators. These are compared against our RAFF module, which employs embedding-specific linear annotators. Table 15 and Table 16 demonstrate that RAFF outperforms all baselines. We attribute this to the inherent semantic diversity across modalities and individual embeddings. Unlike shared mechanisms that fail to capture local nuances, RAFF effectively models these fine-grained sentiment distributions through its embedding-specific approach.

4.8. Hyperparameter Analysis

We conduct extensive single-variable controlled experiments on key hyperparameters to find optimal settings and analyze their impact on performance.

The sequence length of prompts T_p. We investigate the effect of the sequence length of the modality-aware prompts, T_p, on the performance of the PDL model. As shown in Figure 6, with an increase of tokens T_p, the Acc-2 metric exhibits an upward trend. This demonstrates that decoupling sentiment features into multiple tokens, rather than using a single injective embedding, allows for more comprehensive semantics and reduces the loss of important information. Both the CMU-MOSI and CH-SIMS datasets achieve balanced results when $T_p = 8$, while the CMU-MOSEI dataset attains the best comprehensive results when $T_p = 6$.

The depth L of TCFD. We explore how varying the number of TCFD layers influences the model, and present the results in Figure 7. The figure shows that increasing the depth L of the TCFD layer leads to performance improvements to some extent, but deeper layers do not always lead to better performance. A moderately deep TCFD is sufficient to provide adequate learning capacity, whereas excessive depth may lead to overfitting and degrade model performance. The best performance on all three datasets is achieved when $L = 2$.

The threshold σ for dividing the distribution pairs. We examine the optimal threshold value σ for constructing positive and negative distribution pairs in the distribution space, as illustrated in Figure 8. Experimental results indicate that a smaller value of σ, which indicates a stricter partitioning strategy for positive and negative pairs, leads to more effective CDL loss and yields more balanced results. The model achieves balanced performance across the three datasets when $\sigma = 0.1$.

The weight λ for balancing the losses. We evaluate the model's prediction performance under different values of the weighting coefficient λ used in $\mathcal{L}^i$, with the corresponding results shown in Figure 9.

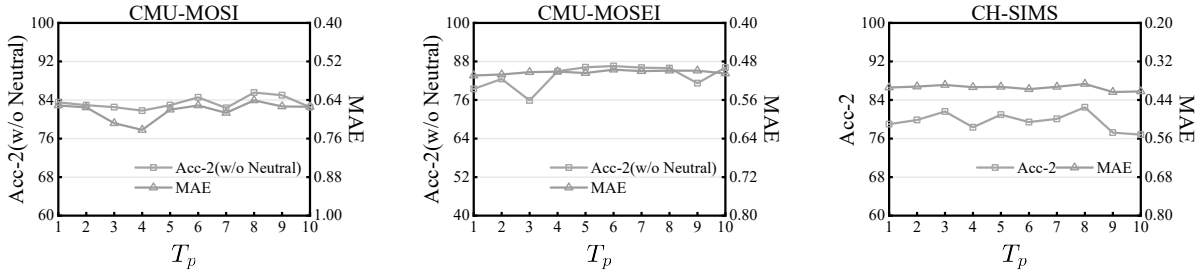

Figure 6: Performance comparison with varying lengths of modality-aware prompt on CMU-MOSI, CMU-MOSEI and CH-SIMS. The experiments are based on the PDL^U model, which utilizes ground-truth unimodal labels to compute the prediction loss and the CDL loss.

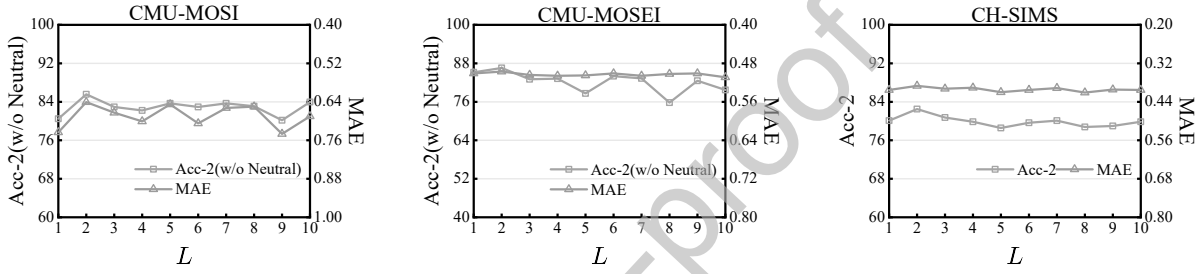

Figure 7: Performance comparison with varying TCFD depths on CMU-MOSI, CMU-MOSEI and CH-SIMS. The experiments are based on the PDL^U model, which utilizes ground-truth unimodal labels to compute the prediction loss and the CDL loss.

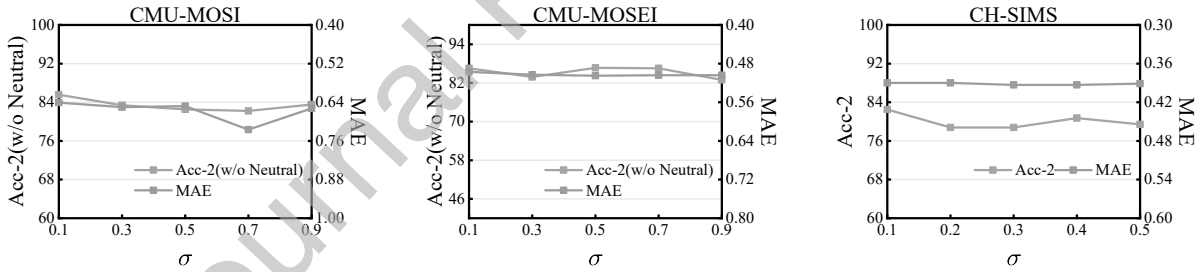

Figure 8: Analysis of the threshold σ on CMU-MOSI, CMU-MOSEI and CH-SIMS. The experiments are based on the PDL^U model, which utilizes ground-truth unimodal labels to compute the prediction loss and the CDL loss.

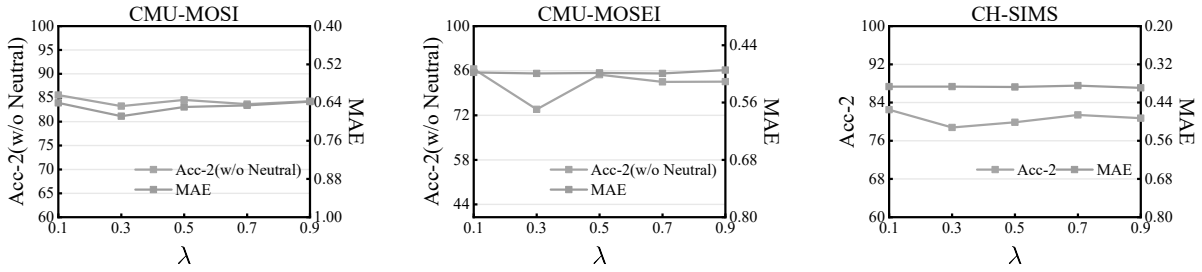

Figure 9: Analysis of the weight λ in $\mathcal{L}^i$ on CMU-MOSI, CMU-MOSEI and CH-SIMS. The experiments are based on the PDL^U model, which utilizes ground-truth unimodal labels to compute the prediction loss and the CDL loss.

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-7$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
GAP	81.34/83.24	81.33/83.28	45.04	0.767	0.777	85.94/87.29	86.02/87.13	55.23	0.500	**0.803**
GMP	83.24/85.52	83.17/85.52	47.23	0.687	0.792	84.16/87.04	84.52/87.02	54.78	0.502	**0.803**
MS-Attn	82.36/84.91	82.12/84.77	45.04	0.692	0.803	86.16/86.16	85.93/85.77	53.62	0.522	0.789
MP-Attn	83.38/85.98	83.16/85.85	46.06	0.697	0.800	85.15/86.74	85.30/86.61	53.10	0.520	0.783
RAFF	**85.57/88.26**	**85.42/88.20**	**49.27**	**0.641**	**0.825**	**86.56/87.34**	**86.59/87.18**	**56.19**	**0.497**	0.801

Table 15: The ablation results of different sequence aggregation strategies on CMU-MOSI and CMU-MOSEI.

Method	Acc-2$_\uparrow$	F1$_\uparrow$	Acc-3$_\uparrow$	Acc-5$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
GAP	78.56	78.92	66.30	44.20	0.404	0.607
GMP	78.34	78.63	67.83	45.08	0.404	0.609
MS-Attn	77.24	77.69	66.08	43.98	0.402	0.608
MP-Attn	81.40	81.38	67.83	44.64	0.401	0.603
RAFF	**82.49**	**82.71**	**69.15**	**46.17**	**0.390**	**0.655**

Table 16: The ablation results of different sequence aggregation strategies on CH-SIMS.

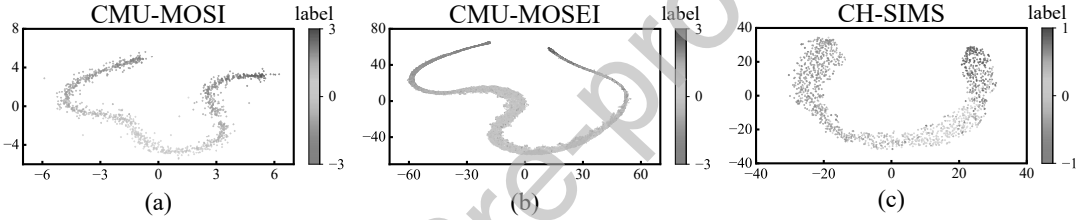

Figure 10: Visualization of the two-dimensional distribution of multimodal features Emb^i using t-SNE on CMU-MOSI, CMU-MOSEI and CH-SIMS, respectively. The color of each point indicates its label y^i. The visualizations are based on PDL^U.

As illustrated, the model performs better when the CDL loss weight λ is smaller. This suggests that while CDL helps align sentiment distributions with labels, its effectiveness is dependent on the guidance provided by the prediction loss $\mathcal{L}^i_{pred}$. We observe that setting λ to 0.1 yields the most balanced results for PDL across different evaluation metrics.

4.9. Computational Cost Analysis

We provide a detailed comparison of computational costs in Table 17. We report the total number of parameters, training duration per epoch, and GFLOPs on CH-SIMS. As shown in the table, compared to the Plain model, the addition of our TCFD, RAFF, and CDL modules in PDL^M and PDL^U results in only a 3.3M increase in parameter count. The PDL^M model shows a strong improvement on Corr while maintaining a nearly identical inference duration and GFLOPs to the Plain baseline. The PDL^U model also achieves a best Corr of 0.655. As the table demonstrates, ALMT is a lightweight model, but its Corr is substantially lower than those of our models. Conversely, MFON is a heavier model, yet its MAE and Corr are inferior to those of our model. We contend that this marginal increase in computational overhead is an acceptable trade-off for the significant improvements in performance and robustness that PDL provides.

4.10. Visualization Analysis

We visualize the two-dimensional spatial distribution of multimodal features and their relationships with labels on the CMU-MOSI, CMU-MOSEI and CH-SIMS datasets, respectively. As shown in Figure 10, the samples exhibit a roughly linear arrangement according to the color gradient. This indicates that our model is capable of capturing the semantic continuity among samples, effectively learning both the intensity and polarity of sentiment semantics embedded in multimodal data.

Model	Parameters$_{\downarrow}$	Duration$_{\downarrow}$	GFLOPs$_{\downarrow}$	MAE$_{\downarrow}$	Corr$_{\uparrow}$
ALMT	**2.6M**	**1.5s**	**3.39**	0.404	0.619
MFON	57.3M	3.7s	4.67	0.420	0.594
Plain	5.9M	2.2s	3.52	**0.390**	0.614
PDL^{M}	9.2M	2.1s	3.40	0.391	0.631
PDL^{U}	9.2M	3.0s	4.06	**0.390**	**0.655**

Table 17: Comparison of parameter size, training duration per epoch, and GFLOPs of PDL on CH-SIMS. Plain: Direct concatenation the "CLS" tokens of unimodal representations without any fusion mechanism, using only the multimodal prediction loss for training. Plain: The plain baseline model that without TCFD, RAFF and CDL module.

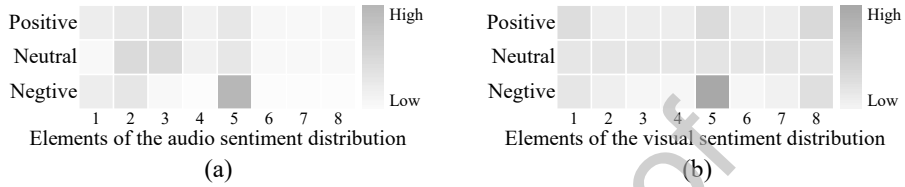

Figure 11: Visualization of semantic distributions based on PDL^{U}. (a) Visualization of the average audio sentiment distribution d_a^i on the training set of the CMU-MOSI dataset. (b) Visualization of the average visual sentiment distribution d_v^i on the training set of the CMU-MOSI dataset.

Method	Conflict(70.90%)			Consist(29.10%)			All		
	Acc-2	F1	MAE	Acc-2	F1	MAE	Acc-2	F1	MAE
Plain	75.30	75.65	0.413	86.46	87.22	0.334	78.56	79.04	0.390
PDL^{M}(Gaussian)	74.07	74.34	0.417	90.22	90.29	0.353	78.77	79.02	0.398
PDL^{M}	75.92	76.06	0.414	90.97	91.31	0.335	80.31	80.52	0.391
PDL^{U}(Gaussian)	76.54	76.49	0.412	87.21	87.47	0.343	79.65	79.67	0.392
PDL^{U}	79.32	79.48	0.426	90.22	90.53	0.299	82.49	82.71	0.390

Table 18: Performance of PDL on the CH-SIMS dataset regarding explicit ambiguity. '↑' indicates better performance as its value increases, '↓' indicates better performance as its value decreases. "PDL^{M}" utilizes multimodal labels as direct substitutes for unimodal labels when computing the unimodal prediction loss and CDL loss. "PDL^{U}" leverages the ground-truth unimodal labels for the unimodal prediction and CDL losses. "Gaussian" refers to the setting where the sentiment semantic distribution is constrained to a Gaussian distribution.

Additionally, we visualize the average sentiment distribution on the training set of the CMU-MOSI dataset. As shown in Figure 11, the semantic activations vary across different sentiments. On the CMU-MOSI dataset, the activation patterns of the audio sentiment distribution and the visual sentiment distribution exhibit a generally consistent trend. This demonstrates that our model can bridge the gap between different modalities and learn a unified cross-modal sentiment representation. Moreover, when comparing sentiment distributions across different sentimental categories, we observe distinct activation patterns. In particular, the 5-th dimension of the distribution is especially sensitive to negative semantics, indicating that the sentiment distribution effectively captures the semantic differences among various sentiment categories.

4.11. MSA Ambiguity Analysis

We perform an in-depth analysis of PDL's effectiveness in handling both explicit and implicit ambiguity. Additionally, we conduct a comparative study where the sentiment semantic distribution is constrained to a Gaussian distribution.

Explicit Ambiguity Analysis. We evaluate PDL on explicit ambiguity, a widely recognized challenge in the field [3, 7, 18], using the CH-SIMS dataset. Based on the alignment between unimodal and multimodal labels, the data is partitioned into consistent and conflicting subsets. As shown in Table 18, the conflicting

Method	High(53.79%)			Low(46.21%)			All		
	Acc-2	F1	MAE	Acc-2	F1	MAE	Acc-2	F1	MAE
Plain	74.25/78.17	73.47/77.54	0.702	93.63/93.63	93.63/93.63	0.729	83.24/85.67	82.81/85.36	0.714
PDL^M(Gaussian)	76.15/78.76	76.07/78.70	0.647	93.38/93.38	93.40/93.40	0.718	84.11/85.82	84.05/85.81	0.681
PDL^M	78.86/81.71	78.81/81.68	0.608	94.64/94.64	94.67/94.67	0.669	86.15/87.96	86.11/87.96	0.636
PDL^U(Gaussian)	75.07/78.76	74.67/78.46	0.625	94.01/94.01	94.01/94.01	0.682	83.82/86.13	83.61/86.00	0.651
PDL^U	76.96/81.42	76.70/81.26	0.628	95.58/95.58	95.59/95.59	0.655	85.57/88.26	85.42/88.20	0.641

Table 19: Performance of PDL on the CMU-MOSI dataset regarding implicit ambiguity. "High" denotes the subset of samples with high label ambiguity, while "Low" represents the subset with low label ambiguity. '↑' indicates better performance as its value increases, '↓' indicates better performance as its value decreases. "PDL^M" utilizes multimodal labels as direct substitutes for unimdoal labels when computing the unimodal prediction loss and CDL loss. "PDL^U" leverages the ground-truth unimodal labels for the unimodal prediction and CDL losses. "Gaussian" refers to the setting where the sentiment semantic distribution is constrained to a Gaussian distribution.

subset accounts for 70.90% of the data, highlighting the inherent non-aligned nature of multimodal sentiment. While performance for all models drops on the conflicting subset due to the heightened explicit ambiguity and semantic complexity, PDL achieves significant improvements on these samples compared to baselines. This confirms PDL's superior robustness in capturing consistent sentiment amidst explicit modal conflicts.

Implicit Ambiguity Analysis. Samples with low absolute sentiment scores often convey ambiguous sentiment, leading to increased annotation uncertainty [7]. Accordingly, we partition the CMU-MOSI test set based on the absolute value of sentiment labels: samples with values greater than 1.5 are categorized as low ambiguity, while the rest are classified as high ambiguity. As shown in Table 19, on the high ambiguity subset, accounting for 53.79% of the data, all models generally perform worse than on the low ambiguity subset. This indicates that samples with neutral sentiments tend to possess higher inherent ambiguity and are potentially more challenging to learn than those with extreme sentiment, which typically exhibit higher certainty. Furthermore, PDL achieves consistent performance gains on both high and low ambiguity subsets, demonstrating its robustness against implicit ambiguity.

Comparison with Gaussian Distribution. As observed in Table 18 and Table 19, this Gaussian constraint leads to a decline in the model's capability to handle ambiguity. This is likely because Gaussian distributions fail to preserve the fine-grained characteristics of decoupled sentiment semantics, given the inherent complexity of multimodal sentiment data.

4.12. Case Study

Figure 12 presents three real-world examples sampled from the CMU-MOSI dataset, covering sentiment extremes: highly positive, neutral, and highly negative cases. These examples demonstrate that our model is capable of accurately predicting sentiment polarity across a range of sentimental intensities.

We also visualize the modality-aware distribution d, which serve as dynamic coefficients for the modality-aware prompts. These weights indirectly reflect the activation of decomposed embeddings across different modalities. In Case 1, the text contains the strongly negative word "bad", while the speaker's vocal tone and facial expressions exhibit mild negativity. However, the facial expressions also display certain positive cues, such as raised eyebrows and a slight smile. Such conflicting signals highlight the presence of both explicit and implicit ambiguities in multimodal sentiment data. Despite these inconsistencies, our model is still able to make an accurate sentiment prediction. Notably, the fifth elements in both d_a and d_v exhibit the highest activation values, indicating the model's effective utilization of informative embeddings from the audio and visual modalities. Case 2 features a speaker delivering a calm and neutral statement, resulting in a balanced distribution across the d vector. In the highly positive Case 3, the speaker exhibits a relaxed and cheerful demeanor, characterized by frequent head movements and nodding. The visual modality in this instance shows higher activation in the first and fifth elements of d_v. These observations suggest that the first and fifth embeddings play significant roles in representing sentiment-related information.

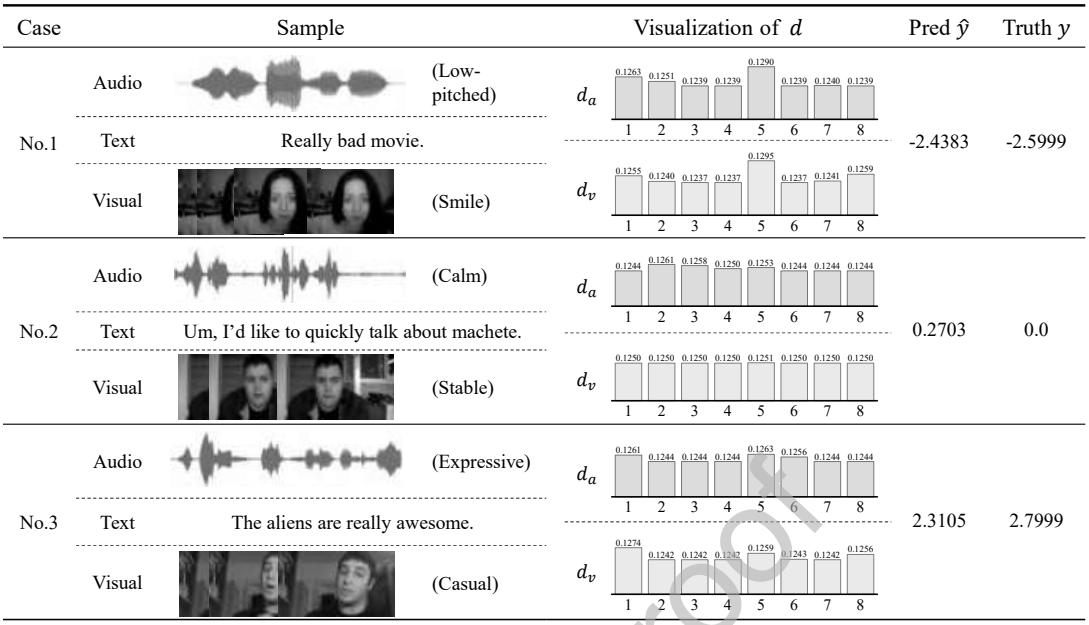

Figure 12: Case study on CMU-MOSI based on PDL^U.

As illustrated, when the audio and visual modalities exhibit similar sentiment polarities, their sentiment distributions tend to align, as demonstrated in Case 3. Conversely, when these modalities convey distinct sentiment polarities, the PDL model can capture distinct unimodal sentiment distributions, as demonstrated in Case 1. By accurately capturing the unimodal sentiments, the PDL model is able to make precise predictions for each sample. As shown, the predicted values closely match the true values, thereby demonstrating the effectiveness of PDL.

4.13. Cross-Dataset and Cross-Lingual Evaluation

To assess PDL's generalization, we conduct cross-dataset and cross-lingual experiments, replacing the monolingual RoBERTa with the multilingual encoders mBERT [43] and XLM-RoBERTa [44], respectively. In this setup, models are trained on one source dataset and evaluated on the other two. As presented in Table 20, our models consistently outperform CASP [42], a method tailored for cross-dataset and cross-lingual learning, on most metrics across both encoder settings, particularly in terms of MAE. These results verify the strong generalization capability and flexibility of PDL in such transfer learning scenarios.

4.14. Comparison with Bayesian Prompt Ensembles.

In contrast to BayesPE, which ensembles predictions to estimate model uncertainty, our approach ensembles decoupled semantics at the feature level to specifically address data ambiguity. Following Bayesian Prompt Ensembles [17], we evaluate three prompt input configurations based on audio-aware prompts (A), visual-aware prompts (V), and the joint application of both (M), aggregating their predictions $\hat{y}$, $Prob_a$ and $Prob_v$ via learned weights. Let BayesPE$(\cdot)$ denote the ensemble strategy optimized on the validation set. For instance, BayesPE($\hat{y}$-A, $\hat{y}$-V) aggregates final predictions $\hat{y}$ generated under separate audio (A) and visual (V) prompt configurations. As evidenced in Table 21, our method outperforms BayesPE-based strategies across the majority of metrics. This demonstrates that aggregating at the feature level effectively preserves complex multimodal sentiment semantics and offers greater flexibility, thereby providing a distinct advantage in modeling multimodal data ambiguity for MSA tasks.

Model	CMU-MOSI(E)→CMU-MOSEI(E)			CMU-MOSEI(E)→CMU-MOSI(E)		
	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$
CASP(2025)	69.12	69.17	0.96	**78.86**	**78.99**	**0.91**
PDL^M(mBERT)	**69.78**	**71.14**	**0.837**	75.66	75.72	0.933
PDL^U(mBERT)	59.97	61.42	0.962	76.53	76.60	0.923
PDL^M(XLM-RoBERTa)	40.16	35.42	1.100	72.01	72.05	1.039
PDL^U(XLM-RoBERTa)	63.68	64.27	0.838	72.26	72.42	1.033
Model	CMU-MOSI(E)→CH-SIMS(C)			CH-SIMS(C)→CMU-MOSI(E)		
	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$
CASP(2025)	51.27	53.15	1.73	48.03	50.43	2.04
PDL^M(mBERT)	67.83	56.67	**0.592**	54.66	52.22	1.365
PDL^U(mBERT)	**69.15**	55.36	0.617	58.16	57.05	1.340
PDL^M(XLM-RoBERTa)	67.61	**67.74**	0.623	**67.35**	**67.36**	**1.213**
PDL^U(XLM-RoBERTa)	63.89	65.29	0.637	59.60	55.08	1.324
Model	CMU-MOSEI(E)→CH-SIMS(C)			CH-SIMS(C)→CMU-MOSEI(E)		
	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$
CASP(2025)	64.23	67.75	1.81	49.09	59.11	1.60
PDL^M(mBERT)	51.86	42.89	0.659	48.72	49.38	0.833
PDL^U(mBERT)	47.92	40.70	0.658	46.62	46.81	0.879
PDL^M(XLM-RoBERTa)	64.55	65.71	**0.474**	62.44	**64.08**	**0.738**
PDL^U(XLM-RoBERTa)	**66.74**	**67.92**	**0.474**	**62.91**	63.27	0.829

Table 20: Cross-lingual and cross-dataset experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS. The model is trained on one source dataset and evaluated on the other two. For CMU-MOSI and CMU-MOSEI, Acc-2 and F1 correspond to the negative/positive (excluding zero) setting. "E" and "C" denote English and Chinese datasets, respectively. PDL^M utilizes multimodal labels as direct substitutes for unimodal labels when computing the unimodal prediction loss and the CDL loss. PDL^U leverages unimodal labels for both the prediction loss and the CDL loss.

5. Conclusion

To address explicit and implicit ambiguities inherent in Multimodal Sentiment Analysis (MSA), we propose a novel divide-and-conquer framework called Prompt-based Distribution Learning (PDL). PDL reorients MSA by focusing on the learning of multimodal sentiment distributions, thereby capturing a broader range of sentiments. Specifically, learnable prompts act as sentiment probes that decompose multimodal features and extract sentiment distributions from each modality. To further enhance multimodal representation learning, we propose a contrastive distribution learning method. Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS validate the effectiveness of PDL, which achieves state-of-the-art or competitive performance. Additional analyses further verify its flexibility and robustness. We believe this approach will inspire greater interest in multimodal representation learning and foster further advancements in the field. For future research, we intend to investigate the application of multimodal self-supervised learning to further enhance distribution learning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62407018), the China Postdoctoral Science Foundation (No. 2023M741305) and the Open Project of Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning (No. 2025AISL004).

Method	CMU-MOSI			CMU-MOSEI			CH-SIMS		
	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$	Acc-2$_\uparrow$	F1$_\uparrow$	MAE$_\downarrow$
PDL^U	**88.26**	**88.20**	**0.641**	87.34	87.18	0.497	**82.49**	**82.71**	0.390
PDL^U+BayesPE($\hat{y}$-A, $\hat{y}$-V)	**88.26**	**88.20**	**0.641**	85.42	85.52	0.507	78.70	79.38	0.402
PDL^U+BayesPE($\hat{y}$-A, $\hat{y}$-V, $\hat{y}$-M)	**88.26**	**88.20**	**0.641**	86.32	86.32	0.502	79.21	79.75	0.400
PDL^U+BayesPE($Prob_a$-A, $Prob_v$-V)	86.89	86.88	0.650	**87.70**	**87.74**	**0.493**	80.30	80.64	**0.385**

Table 21: Results of PDL combined with Bayesian Prompt Ensembles. A: solely audio-aware prompts; V: solely visual-aware prompts; M: joint input of both audio- and visual-aware prompts. BayesPE(·) denotes the optimization process on the validation set. E.g., BayesPE($\hat{y}$-A, $\hat{y}$-V) aggregates final predictions derived independently from A and V prompts.

References

[1] U. Singh, K. Abhishek, H. K. Azad, A Survey of Cutting-edge Multimodal Sentiment Analysis, ACM Comput. Surv. 56 (9), ISSN 0360-0300.

[2] F. Rodrigues, F. Pereira, Deep learning from crowds, in: Proceedings of the AAAI conference on artificial intelligence, vol. 32, 2018.

[3] D. Hazarika, R. Zimmermann, S. Poria, MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 1122–1131, 2020.

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

[5] J. Yang, Y. Yu, D. Niu, W. Guo, Y. Xu, ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 7617–7630, 2023.

[6] C. Fan, J. Lin, R. Mao, E. Cambria, Fusing pairwise modalities for emotion recognition in conversations, Information Fusion 106 (2024) 102306.

[7] C. Tang, T. Shen, X. Gong, C. Zhao, T. Zhang, DFMU: Distribution-based Framework for Modeling Aleatoric Uncertainty in Multimodal Sentiment Analysis, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 8250–8258, 2025.

[8] L. Xiao, R. Mao, S. Zhao, Q. Lin, Y. Jia, L. He, E. Cambria, Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis, IEEE Transactions on Affective Computing .

[9] M. Poesio, R. Artstein, The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account, in: A. Meyers (Ed.), Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, 76–83, 2005.

[10] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, M. Valstar, COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2) (2024) 805–822.

[11] Y. Wu, C. Wang, J. Li, W. Zhang, X. Jiang, Uncertainty-Aware Gradient Modulation and Feature Masking for Multimodal Sentiment Analysis, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 321–335, 2024.

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

[13] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2236–2246, 2018.

[14] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, T. Yu, Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis, arXiv preprint arXiv:2310.05804 .

[15] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognition 136 (2023) 109259.

[16] Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, H. T. Shen, Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 26866–26875, 2024.

[17] F. Tonolini, N. Aletras, J. Massiah, G. Kazai, Bayesian Prompt Ensembles: Model Uncertainty Estimation for Black-Box Large Language Models, in: Findings of the Association for Computational Linguistics: ACL 2024, 12229–12272, 2024.

[18] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI conference on artificial intelligence, vol. 35, 10790–10797, 2021.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, 8748–8763, 2021.

[20] R. Mao, Q. Liu, K. He, W. Li, E. Cambria, The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection, IEEE Transactions on Affective Computing 14 (3) (2023) 1743–1753.

[21] R. Mao, K. He, C. Ong, Q. Liu, E. Cambria, MetaPro 2.0: Computational Metaphor Processing on the Effectiveness of Anomalous Language Modeling, in: Findings of the Association for Computational Linguistics, 9891–9908, 2024.
[22] Y. Wu, C. Wang, J. Li, W. Zhang, X. Jiang, Uncertainty-Aware Gradient Modulation and Feature Masking for Multimodal Sentiment Analysis, in: Pattern Recognition and Computer Vision: 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18–20, 2024, Proceedings, Part XI, 321–335, 2024.
[23] J. Li, C. Wang, Z. Luo, Y. Wu, X. Jiang, Modality-Dependent Sentiments Exploring for Multi-Modal Sentiment Classification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7930–7934, 2024.
[24] X. Geng, Label distribution learning, IEEE Transactions on Knowledge and Data Engineering 28 (7) (2016) 1734–1748.
[25] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, X. Geng, Deep label distribution learning for apparent age estimation, in: Proceedings of the IEEE international conference on computer vision workshops, 102–108, 2015.
[26] N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le, A. Nguyen, Uncertainty-aware label distribution learning for facial expression recognition, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 6088–6097, 2023.
[27] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance as a metric for image retrieval, International journal of computer vision 40 (2000) 99–121.
[28] Y. Wu, Z. Lin, Y. Zhao, B. Qin, L.-N. Zhu, A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis, in: Findings of the association for computational linguistics: ACL-IJCNLP 2021, 4730–4738, 2021.
[29] S. Zhang, J. Zhang, Z. Zhang, L. Li, Multimodal Mixture of Low-Rank Experts for Sentiment Analysis and Emotion Recognition, arXiv preprint arXiv:2505.14143 .
[30] J. Lin, Y. Wang, Y. Xu, Q. Liu, Semi-IIN: Semi-supervised Intra-inter modal Interaction Learning Network for Multimodal Sentiment Analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 1411–1419, 2025.
[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, ArXiv abs/1907.11692.
[32] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python., in: SciPy, 18–24, 2015.
[33] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency, OpenFace 2.0: Facial Behavior Analysis Toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 59–66, 2018.
[34] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, CoRR abs/1412.6980.
[35] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognition 136 (2023) 109259.
[36] Y. Wang, H. Jian, J. Zhuang, H. Guo, Y. Leng, SSLMM: Semi-Supervised Learning with Missing Modalities for Multimodal Sentiment Analysis, Information Fusion 120 (2025) 103058.
[37] S. Wu, D. He, X. Wang, L. Wang, J. Dang, Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 1601–1609, 2025.
[38] H. Shi, Y. Pu, Z. Zhao, J. Huang, D. Zhou, D. Xu, J. Cao, Co-space Representation Interaction Network for multimodal sentiment analysis, Knowledge-Based Systems 283 (2024) 111149.
[39] C. Fan, K. Zhu, J. Tao, G. Yi, J. Xue, Z. Lv, Multi-level Contrastive Learning: Hierarchical Alleviation of Heterogeneity in Multimodal Sentiment Analysis, IEEE Transactions on Affective Computing (2024) 1–17.
[40] X. Zhang, W. Wei, S. Zou, Modal Feature Optimization Network with Prompt for Multimodal Sentiment Analysis, in: Proceedings of the 31st International Conference on Computational Linguistics, 4611–4621, 2025.
[41] L. Zhu, H. Zhao, Z. Zhu, C. Zhang, X. Kong, Multimodal sentiment analysis with unimodal label generation and modality decomposition, Information Fusion 116 (2025) 102787.
[42] Z. Guo, T. Jin, W. Xu, W. Lin, Y. Wu, Bridging the gap for test-time multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 16987–16995, 2025.
[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186, 2019.
[44] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 8440–8451, 2020.

Declaration of interests

Chengji Wang reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.