

PaSE: Prototype-aligned Calibration and Shapley-based Equilibrium for Multimodal Sentiment Analysis

Kang He^{1,2}, Boyu Chen¹, Yuzhe Ding¹, Fei Li¹, Chong Teng¹, Donghong Ji^{1*}

¹ Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University

² Shanghai Innovation Institute

{hegang0225, chenboyu1225, yuzheding, lifei_csnlp, chongteng, dhji}@whu.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) seeks to understand human emotions by integrating textual, acoustic, and visual signals. Although multimodal fusion is designed to leverage cross-modal complementarity, real-world scenarios often exhibit *modality competition*: dominant modalities tend to overshadow weaker ones, leading to *suboptimal performance*. In this paper, we propose PaSE, a novel **Prototype-aligned Calibration and Shapley-optimized Equilibrium** framework, which enhances collaboration while explicitly mitigating modality competition. PaSE first applies Prototype-guided Calibration Learning (PCL) to refine unimodal representations and align them through an Entropic Optimal Transport mechanism that ensures semantic consistency. To further stabilize optimization, we introduce a Dual-Phase Optimization strategy. A prototype-gated fusion module is first used to extract shared representations, followed by Shapley-based Gradient Modulation (SGM), which adaptively adjusts gradients according to the contribution of each modality. Extensive experiments on IEMOCAP, MOSI, and MOSEI confirm that PaSE achieves the superior performance and effectively alleviates modality competition.

Introduction

Multimodal Sentiment Analysis (MSA) aims to infer human emotions by integrating information from diverse modalities such as text, audio, and visual data. Compared to traditional unimodal approaches (Fan et al. 2022; Wang et al. 2024), MSA leverages the complementary nature of modalities to achieve deeper emotional understanding (Poria et al. 2016; Zhu et al. 2024; Xu, Jiang, and Liang 2024). With the rapid progress in representation learning (He et al. 2025a,b) and the proliferation of multi-source data, MSA has become a research hotspot, finding applications in dialogue systems (Zhang et al. 2024), stance detection (Ding et al. 2025) and social media analysis (Zhai et al. 2024).

A central challenge in MSA lies in effectively fusing heterogeneous signals to optimize cross-modal interactions. Early works focused on feature-level concatenation (Poria et al. 2016; Yu and Jiang 2019), while subsequent efforts explored decision-level fusion by aggregating predictions from individual modalities (Zhu et al. 2024; Zhuang et al.

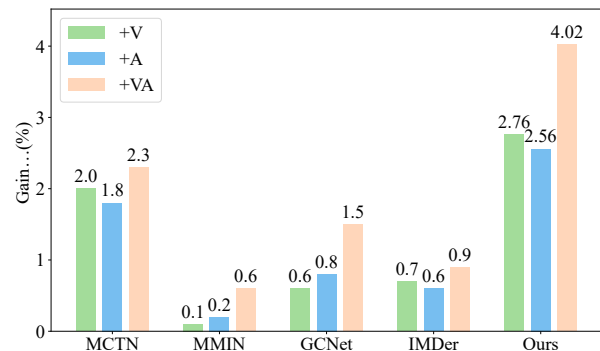


Figure 1: Performance improvements (F1 score) from adding audio (+A), visual (+V), or both modalities (+VA) to text-only baselines on the CMU-MOSI datasets.

2024). These paradigms generally presume that *modality collaboration dominates the fusion process* and that combining modalities will inherently yield performance gains.

However, real-world MSA tasks often involve a *dynamic trade-off between modalities*. Variations in data distribution, representational capacity, convergence rate, and noise levels can result in imbalanced learning across modalities (Wang, Tran, and Feiszli 2020; Fujimori et al. 2020; Huang et al. 2022). Such imbalance may lead to suboptimal fusion outcomes, where the benefits of multimodal integration do not meet expectations (Zhou et al. 2023). As a result, the ideal of **cross-modal Pareto optimality** is rarely realized in practice.

Huang et al. (2022) observed that unimodal models can occasionally outperform their multimodal counterparts—an effect termed *modality competition*. As shown in Figure 1, incorporating audio and visual modalities into a text-only baseline yields limited performance gains, suggesting that modality interactions may involve interference or rivalry rather than straightforward cooperation. Recent studies (Wu et al. 2022; Fan et al. 2023; Zhou et al. 2023) further characterize multimodal learning as a non-cooperative game, wherein dominant modalities exert greater influence by transmitting stronger gradient signals (Peng et al. 2022), suppressing the contributions of weaker modalities. While this phenomenon has garnered growing attention, the intricate dynamics between modality collaboration and competition remain under-explored, posing a persistent challenge in MSA. These issues

*Corresponding author

underscore the need for more adaptive fusion strategies that explicitly address modality imbalance and competition while promoting complementary integration.

To address the challenges of modality competition, we propose PaSE, a prototype-guided fusion framework that promotes cross-modal synergy while mitigating dominance effects. PaSE first calibrates unimodal features via Prototype-guided Calibration Learning (PCL), then aligns cross-modal semantics through Entropic Optimal Transport-based alignment learning (CAL). To further ensure balanced learning, we introduce a Dual-Phase Optimization strategy. In the first phase, Prototype-Gated Fusion mechanism (PGF) integrates modalities to learn a robust overall representation, allowing dominant modalities to guide early training. In the second phase, Shapley-based Gradient Modulation (SGM) dynamically reweights gradients based on each modality’s marginal contribution, mitigating long-term imbalance. Our main contributions are summarized as follows:

1. We propose PaSE, a multimodal framework that integrates prototype-aligned calibration with Shapley-optimized equilibrium to address modality competition in MSA.
2. We introduce an entropic optimal transport-based cross-modal prototype alignment mechanism that enforces structural consistency and bidirectional matching, effectively mitigating semantic discrepancies across modalities.
3. We develop a dynamic fusion mechanism with Shapley-based gradient modulation to adaptively balance modality contributions and improve collaborative optimization.
4. Extensive experiments on three benchmark datasets (IEMOCAP, MOSI, and MOSEI) demonstrate that PaSE achieves state-of-the-art performance and effectively mitigates modality competition.

Related Work

Multimodal Sentiment Analysis

Fusion is a central challenge in MSA (Yang, Xu, and Gao 2020; Mai et al. 2022). Early studies concatenate features from multiple modalities for joint representation learning (Poria et al. 2016; Yu and Jiang 2019), while later approaches adopt decision-level fusion by aggregating unimodal predictions (Zhu et al. 2024; Zhuang et al. 2024). With the rise of attention mechanisms, cross-modal attention has become popular for adaptive fusion (Tsai et al. 2019; Yang, Xu, and Gao 2020; Mai et al. 2022). Another line of work explores modality disentanglement. MISA (Hazarika, Zimmermann, and Poria 2020) separates shared and private representations via dedicated encoders. Follow-up methods incorporate this idea with distillation (Li, Wang, and Cui 2023) and contrastive learning (Yang et al. 2023) to boost cross-modal interaction. Despite these advances, most methods neglect the adverse effects of modality competition. To address this, we propose a prototype-guided calibration alignment and a Shapley-based gradient modulation mechanism to balance inter-modal collaboration and competition.

Modality Competition

Modality imbalance (Fernando et al. 2021; Peng et al. 2022; Fan et al. 2023) arises from differences in information quality

and data distribution, often causing dominant modalities to suppress weaker ones during training. Prior solutions include gradient modulation (Peng et al. 2022; Fernando et al. 2021), adaptive learning rates (Wang, Tran, and Feiszli 2020; Wu et al. 2022), and modality-specific objectives (Fan et al. 2023, 2024b). However, most rely on indirect signals such as gradient norms or losses, lacking principled quantification of each modality’s contribution. We address this by introducing Shapley values to estimate marginal utility and guide gradient modulation, enabling more effective mitigation of modality competition.

Task Setup

MSA aims to predict the emotional polarity (e.g., positive, neutral, negative) or emotional intensity of a given utterance by integrating information from multiple modalities, such as text (t), audio (a), and vision (v). Formally, for each modality, we represent the sequence of each modality as $H^m \in \mathcal{R}^{B_m \times d_m}$, where $m \in \mathcal{M} = \{t, a, v\}$ represents different modalities, B_m and d_m denote the corresponding dimensions and sequence lengths.

Methodology

As illustrated in Figure 2, PaSE consists of three main modules: namely the intra-modal prototype calibration learning (PCL), the cross-modal alignment learning (CAL), and the dual-phase optimization module. The latter is further divided into the modality fusion module and the Shapley-based gradient modulation (SGM). Detailed descriptions are provided in the following subsections.

Intra-modality Prototype Calibration (PCL)

As previously noted, imbalanced expressive power across modalities often causes weaker modalities to be overshadowed by dominant ones (Wu et al. 2022; Fan et al. 2023), hindering effective fusion and degrading overall performance (Zhou et al. 2023). To mitigate this, we introduce a Prototype-guided Calibration Learning (PCL) strategy. The key idea is to leverage category prototypes to guide weaker modalities toward a more discriminative and structured feature space, thereby avoiding suppression by stronger modalities. This also facilitates better modeling of both intra-class cohesion and inter-class separability within each modality. Each category prototype is defined as the centroid of features belonging to that category in the corresponding modality:

$$c_k^m = \frac{1}{|B_k|} \sum_{i \in B_k} h_i^m \quad (1)$$

where c_k^m is the prototype of category k in modality m , B_k is the set of samples belonging to category k in the mini-batch. h_i^m is the feature representation of sample i in modality m . To ensure stability during prototype updates, we adopt a momentum-based moving average strategy (He et al. 2020):

$$c_k^m \leftarrow \gamma c_k^m + (1 - \gamma) \frac{1}{|B_k|} \sum_{i \in B_k} h_i^m \quad (2)$$

To encourage samples within the same category to form a compact and consistent distribution in each modality, we

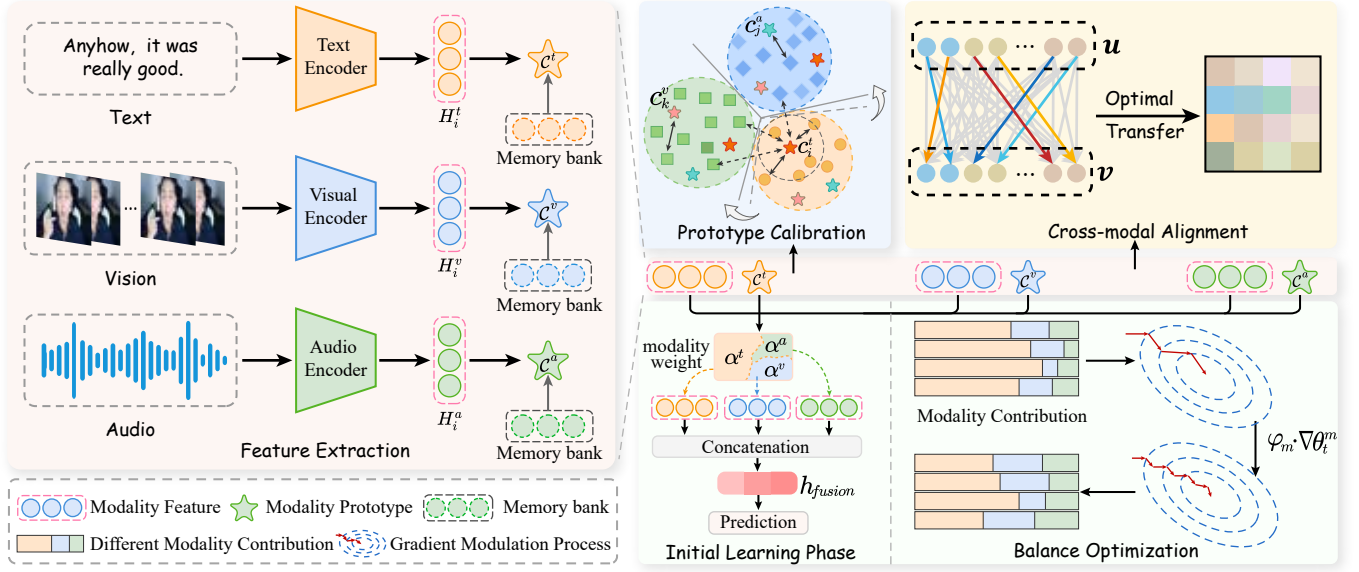


Figure 2: The overall architecture of our proposed model PaSE.

introduce the intra-modal prototype calibration loss \mathcal{L}_{intra}^m :

$$\mathcal{L}_{intra}^m = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{(\phi(h_i^m, c_{y_i}^m)/\tau)}}{\sum_{k=1}^K e^{(\phi(h_i^m, c_k^m)/\tau)}} \quad (3)$$

where $c_{y_i}^m$ is the prototype vector for the category of sample i in modality m , $\phi(h_i^m, c_k^m)$ is the similarity function, τ is the temperature parameter, and N is the batch size. This loss promotes tighter intra-class clustering within each modality, allowing the model to better utilize the expressive capacity of weaker modalities. Unlike approaches that rely on dominant modalities to shape representations, our method encourages each modality to refine its features based on its own semantic structure, thus alleviating modality imbalance.

Cross-modal Alignment Learning via Entropic Optimal Transport (CAL)

Intra-modal optimization alone is insufficient to guarantee robust cross-modal fusion, particularly due to the inherent distributional heterogeneity across modalities (Yang et al. 2023). To address this, we align different modalities using an Entropic Optimal Transport (EOT) framework.

Formally, let $\{c_k^{(m)}\}_{k=1}^K$ and $\{c_l^{(n)}\}_{l=1}^K$ denote the sets of class prototypes from modalities m and n , respectively. These prototypes are treated as discrete distributions $P^{(m)}$ and $P^{(n)}$ in the latent space. The alignment objective is to minimize the transportation cost between the two distributions, formulated as the Wasserstein distance:

$$\mathcal{L}_{proto}^{m,n} = \inf_{\pi \in \Pi(P^m, P^n)} \mathbb{E}_{(x,y) \sim \pi} [\|x - y\|_2^2] \quad (4)$$

where $\Pi(P^m, P^n)$ denotes the set of all joint distributions (transport plans) with marginals P^m and P^n , and π defines the matching strategy between prototype pairs across modalities. However, directly optimizing Eq. 4 is often computationally prohibitive (Gushchin et al. 2023; Kassraie et al.

2024). To this end, we adopt an entropy-regularized relaxation, known as Entropic Optimal Transport (EOT), which enables stable and tractable alignment. Given a ground cost matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$ with entries $C_{kl} = d(c_k^{(m)}, c_l^{(n)})$, the EOT objective is defined as:

$$\min_{\mathbf{Q} \in \mathcal{U}(\mathbf{u}, \mathbf{v})} \langle \mathbf{Q}, \mathbf{C} \rangle_F + \lambda \mathcal{H}(\mathbf{Q}), \quad \text{s.t. } \mathbf{Q} \mathbf{1}_K = \mathbf{u}, \mathbf{Q}^\top \mathbf{1}_K = \mathbf{v} \quad (5)$$

where \mathbf{Q} is the transport matrix, $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, and $\mathcal{H}(\mathbf{Q}) = -\sum_{k,l} \mathbf{Q}_{kl} \log \mathbf{Q}_{kl}$ is the entropy regularizer. The marginal constraints \mathbf{u} and \mathbf{v} are typically set to uniform distributions, and $\mathcal{U}(\mathbf{u}, \mathbf{v})$ denotes the space of doubly stochastic matrices. To encourage bidirectional semantic consistency, we compute the transport plans in both directions ($\mathbf{Q}^{(m \rightarrow n)}, \mathbf{Q}^{(n \rightarrow m)}$) and define a symmetric matching loss:

$$\mathcal{L}_{match} = \frac{1}{2} \left(\langle \mathbf{Q}^{(m \rightarrow n)}, \mathbf{C} \rangle_F + \langle \mathbf{Q}^{(n \rightarrow m)}, \mathbf{C}^\top \rangle_F \right) \quad (6)$$

which promotes mutual alignment between modality pairs and alleviates one-sided matching bias. Furthermore, to reinforce the consistency between forward and backward alignment, we introduce a consistency regularization that penalizes the discrepancy between the two transport matrices:

$$\mathcal{L}_{reg} = \|\mathbf{Q}^{(m \rightarrow n)} - (\mathbf{Q}^{(n \rightarrow m)})^\top\|_F^2 \quad (7)$$

Lastly, we incorporate a structure-preserving regularizer to encourage identity-like transport matrices, thereby preserving intra-class semantic integrity:

$$\Omega(\mathbf{Q}) = \|\mathbf{Q} - \mathbf{I}_K\|_F^2 \quad (8)$$

where \mathbf{I}_K denotes the identity matrix. By integrating the above components, the total loss function for cross-modal prototype alignment is formulated as:

$$\mathcal{L}_{inter} = \sum_{(m,n) \in \mathcal{P}} \left(\mathcal{L}_{match}^{(m,n)} + \alpha \mathcal{L}_{reg}^{(m,n)} + \beta \Omega(\mathbf{Q}) \right) \quad (9)$$

where the hyperparameters α and β control the relative importance of the consistency constraint and the structure-preserving regularization.

Dual-Phase Optimization Module

Modality Initialization Learning Phase Early-stage multimodal learning often suffers from modality dominance, where stronger modalities overshadow weaker ones (Yu and Jiang 2019; Yang, Xu, and Gao 2020; Mai et al. 2022). To mitigate this issue, we introduce an entropy-guided modality weighting strategy, which assigns importance to each modality based on the uncertainty of its predictions:

$$E(H^m) = 1 - \frac{1}{N} \sum_{i=1}^N p(y_i|h_i^m) \log p(y_i|h_i^m) \quad (10)$$

The resulting modality weight is computed via a softmax over all modalities:

$$\alpha^{(m)} = \frac{\exp(E(H^m))}{\sum_{k \in \mathcal{M}} \exp(E(H^k))} \quad (11)$$

Modality Fusion We employ a Prototype-Gated Fusion (PGF) module to dynamically integrate modality-specific representations. Each modality’s representation h^m is modulated via a gating mechanism conditioned on both its learned embedding and corresponding class prototype and obtain the fused representation:

$$g^{(m)} = \sigma(W_g^m[\alpha^{(m)}h^m \oplus c_y^m]) \quad (12)$$

$$h_{\text{fusion}} = \text{FFN} \left(\bigoplus_{m \in \mathcal{M}} g^{(m)} \cdot h^m \right) \quad (13)$$

The final sentiment prediction \hat{y} is computed by:

$$\hat{y} = \text{Softmax}(\text{MLP}(h_{\text{fusion}})) \quad (14)$$

Objective Function The training objective integrates both the classification loss and prototype alignment regularization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mu \mathcal{L}_{\text{align}} \quad (15)$$

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

$$\mathcal{L}_{\text{align}} = \sum_{m \in \mathcal{M}} \alpha^{(m)} \mathcal{L}_{\text{intra}}^m + \mathcal{L}_{\text{inter}} \quad (17)$$

This function allows the dominant modality to guide training while improving weaker ones through alignment feedback.

Modality Balance Optimization Phase To address the long-term imbalance in gradient updates caused by modality dominance, we propose a Shapley-based Gradient Modulation (SGM) mechanism. Inspired by cooperative game theory (Gul 1989), we quantify the marginal contribution of each modality using the Shapley value:

$$\psi_m(u) = \sum_{S \subseteq \mathcal{M} \setminus \{m\}} \frac{|S|!(k - |S| - 1)!}{k!} [u(S \cup \{m\}) - u(S)] \quad (18)$$

The utility function $u(S)$ is defined as:

$$u(S) = \frac{\rho}{1 + \mathcal{L}_{\text{inter}}^S} + \frac{1 - \rho}{1 + \mathcal{L}_{\text{intra}}^S} \quad (19)$$

where $\rho \in [0, 1]$ controls the balance between inter- and intra-modal performance based on Eq. 3 and Eq. 9. We normalize Shapley values and use them to modulate gradient updates:

$$\tilde{\psi}_m = \frac{\psi_m}{\sum_k \psi_k}, \quad \varphi_m = \exp \left(\frac{\tilde{\psi}_{\min}}{\tilde{\psi}_m} - 1 \right) \quad (20)$$

where φ_m serves as a modulation factor that dynamically adjusts the learning rate of each modality by amplifying the updates of weaker modalities and suppressing the perturbations of dominant ones. The parameter update of modality m at iteration $t + 1$ is:

$$\theta_{t+1}^m = \theta_t^m - \eta \cdot \varphi_m \cdot \frac{\partial \mathcal{L}^{(t)}}{\partial \theta^m} \quad (21)$$

This stage encourages optimization focus to shift toward under-performing modalities, thereby promoting a balanced fusion process.

Experiments

Datasets and Evaluation Metrics

To validate the effectiveness of PaSE, we conduct experiments on three benchmark datasets: CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Bagher Zadeh et al. 2018), and IEMOCAP (Busso et al. 2008).

We report the averaged results over five runs with different random seeds. For classification, we use 2-class (Acc-2) and 7-class (Acc-7) accuracy for MOSI and MOSEI, and F1 for IEMOCAP. For regression, we report Mean Absolute Error (MAE) and correlation (Corr), where all metrics except MAE favor higher values. Note that Acc-2 and F1 on CMU-MOSI and CMU-MOSEI are reported in two forms (denoted as ‘-/-’): one includes zero as negative/non-negative, the other excludes it as negative/positive.

Baselines

In this study, we select a diverse set of current SOTA baselines to conduct a comprehensive comparison. These methods include MuLT (Tsai et al. 2019), MAG-BERT (Rahman et al. 2020), SelfMM (Yu et al. 2021), HyCon (Mai et al. 2022), ConKI (Yu et al. 2023), ConFEDE (Yang et al. 2023), CLGSI (Yang, Dong, and Qiang 2024), MCL-MCF (Fan et al. 2024a), MFON (Zhang, Wei, and Zou 2025), GLoMo (Zhuang et al. 2024), EUAR (Gao et al. 2024), KEBR (Zhu et al. 2024), Semi-IIN (Lin et al. 2025), and MSamba (He et al. 2025c).

Implementation Details

To ensure fair comparison with existing SOTA methods, we adopt the standard feature extractors: Facet for visual, COVAREP (Degottex et al. 2014) for acoustic, and BERT_{base} (Devlin et al. 2019) for textual modalities on CMU-MOSI and CMU-MOSEI. We use Adam optimizer with a learning rate of 1e-5, a batch size of 64, and train for 200 epochs. We

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2(↑)	F1(↑)	Acc-7(↑)	Corr(↑)	MAE(↓)	Acc-2(↑)	F1(↑)	Acc-7(↑)	Corr(↑)	MAE(↓)
MuLT [♡] _{ACL19}	79.51/80.47	79.46/80.49	36.74	0.667	0.892	81.10/83.63	81.05/83.46	52.34	0.605	0.671
MAG-BERT [◇] _{ACL20}	82.37/84.43	82.50/84.61	43.62	0.727	0.781	82.51/84.82	82.77/84.71	52.67	0.755	0.543
SelfMM [◇] _{AAAI21}	82.54/84.77	84.42/85.95	45.79	0.795	0.712	82.68/84.96	82.95/84.93	53.46	0.767	0.529
HyCon [◇] _{TAC22}	-/85.20	82.68/84.91	46.60	0.790	0.713	-/85.40	-/85.60	52.80	0.776	0.601
ConKI [◇] _{ACL23}	84.37/86.13	84.33/86.13	48.43	0.816	<u>0.681</u>	82.73/86.25	83.08/86.15	54.25	0.782	0.529
ConFEDE [◇] _{ACL23}	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	54.86	0.780	0.522
CLGSI [◇] _{NAACL24}	83.97/86.43	83.63/86.25	47.96	0.790	0.703	84.01/86.32	84.21/86.18	54.56	0.763	0.532
MCL-MCF [◇] _{TAC24}	84.90/87.30	84.70/87.20	-	0.799	0.692	84.20/86.40	84.40/86.30	-	0.767	0.536
MFON [◇] _{COLING25}	84.84/86.89	84.75/86.86	44.90	0.797	0.725	82.70/86.32	83.13/86.29	53.72	0.780	0.528
GLoMo [◇] _{MM25}	84.10/86.70	83.90/86.60	48.30	0.782	0.718	83.70/86.50	84.00/86.40	55.00	0.771	0.539
EUAR [*] _{MM25}	83.24/84.58	84.22/84.58	48.18	0.789	0.711	82.20/85.64	82.02/85.46	51.07	0.785	0.596
KEBR [*] _{MM25}	84.84/87.27	84.83/87.25	47.81	0.819	0.683	84.01/86.74	84.25/86.68	54.37	0.799	0.517
Semi-IIN [*] _{AAAI25}	85.28/87.04	85.19/87.00	46.50	<u>0.822</u>	0.679	84.98/87.70	85.27/87.65	55.89	<u>0.804</u>	0.497
MSAmba [*] _{AAAI25}	<u>85.99/87.43</u>	<u>85.99/87.40</u>	<u>49.67</u>	0.809	0.707	<u>85.78/86.86</u>	<u>85.99/86.93</u>	54.21	0.796	<u>0.507</u>
PaSE(ours)	86.40/88.32	86.34/88.25	50.92	0.847	0.695	86.07/88.10	86.21/87.96	<u>55.76</u>	0.831	0.523

Table 1: Comparison with SOTA methods on MOSI and MOSEI. Bold indicates the best result, and underlining denotes the second-best. Note: ♡ and ◇ represents results obtained from Zhuang et al. (2024) and Yu et al. (2023), respectively. “-” denotes the result is not provided, * indicates reproduced results from the public code. Other results taken directly from the original paper.

Method	Happy	Sad	Angry	Neutral
MCTN	83.1	82.8	84.6	67.7
Self-MM	90.8	86.7	88.4	<u>72.7</u>
SMIL	86.8	85.2	84.9	68.9
GCNet	87.7	86.9	85.2	71.1
UMDF	87.9	86.5	85.8	70.5
DMD	<u>91.1</u>	<u>88.4</u>	<u>88.6</u>	72.2
CorrKD	87.5	85.9	86.1	71.5
PaSE (ours)	91.5	88.6	89.4	73.2

Table 2: Comparison on IEMOCAP dataset across different emotion classes (F1-score, ↑).

set $\gamma = 0.98$ for prototype updates, with $\lambda = 0.01$, $\mu = 0.1$, and alignment-enhancing factors $\alpha = 0.1$, $\beta = 0.05$ to improve cross-modal consistency. All experiments are conducted on a NVIDIA A100 GPU. A warm-up phase is first employed to stabilize modality-specific learning and prevent early overfitting. Training then automatically transitions to the Shapley-guided balancing phase once the validation entropy stabilizes, with validation accuracy monitored to ensure performance remains stable.

Results and Analysis

Comparison with SOTA Methods

As shown in Table 1 and Table 2, we conduct extensive evaluations on three widely-used benchmarks: CMU-MOSI, CMU-MOSEI, and IEMOCAP. Our proposed PaSE, consistently outperforms existing SOTA approaches across all datasets and most evaluation metrics. Notably, on CMU-MOSI, PaSE surpasses the most recent SOTA model, MSAmba, achieving relative gains of 0.89% in Acc-2 and 1.25% in Acc-7. Comparable consistent gains are also observed on CMU-MOSEI, highlighting the effectiveness of the prototype-guided alignment mechanism in capturing modality-consistent semantics and mitigating modality noise.

Model	CMU-MOSI		CMU-MOSEI		IEMOCAP
	Acc-2(↑)	F1(↑)	Acc-2(↑)	F1(↑)	WA(↑)
<i>Ablation Study</i>					
PaSE	86.40/88.32	86.34/88.25	86.07/88.10	86.21/87.96	80.47
- PCL	85.21/87.25	85.17/87.16	84.73/86.91	84.96/86.83	78.54
- CAL	85.04/86.80	85.06/86.82	84.54/86.69	84.47/86.62	78.13
- SGM	83.65/85.47	83.60/85.39	83.13/85.03	83.46/85.08	76.80
<i>Ablation on Fusion Strategy</i>					
- PGF	84.57/86.37	82.48/86.32	83.06/86.05	84.45/85.90	77.94
+ Sum	84.45/86.92	84.34/86.26	83.73/86.59	83.89/86.53	78.43
+ Con	85.92/87.40	85.85/87.37	85.37/87.14	85.51/87.10	79.16
+ Att	85.60/87.26	85.53/87.19	85.25/86.85	85.40/86.97	78.75

Table 3: Ablation studies for each module in our proposed model PaSE on MOSI, MOSEI and IEMOCAP datasets. We adopt weighted accuracy (WA) as evaluation metrics for IEMOCAP datasets.

On the IEMOCAP dataset, PaSE achieves the highest performance across all four emotion categories, with an average improvement of 2.93% over prior methods. This substantial performance gain reflects the superior discriminative capability and resilience of our model in more complex emotion recognition settings. We attribute these improvements to the explicit alignment of modality-specific features toward shared semantic prototypes, which enhances intra-class compactness and inter-class separability in the fused representation space.

Ablation Study

To evaluate the contribution of each component in PaSE, we conduct a series of ablation experiments on the MOSI and MOSEI datasets. As shown in Table 3, removing the prototype-guided calibration (PCL) leads to a clear drop in performance, with Acc-2 decreasing by 1.07% on MOSI and 1.19% on MOSEI. This highlights the importance of the prototype-based structure in enforcing intra-class compactness and inter-class separability within modalities, which are

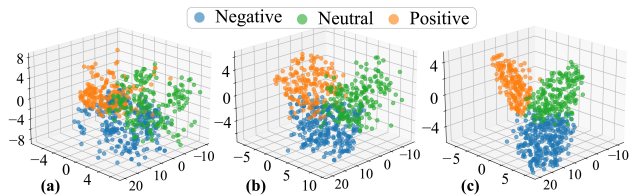


Figure 3: t-SNE visualization of fused feature representations obtained using different models on MOSI test set. (a) SelfMM, (b)EUAR, and (c) PaSE (ours).

Model	CMU-MOSI		CMU-MOSEI		IEMOCAP
	Acc-2(↑)	F1(↑)	Acc-2(↑)	F1(↑)	WA(↑)
A	60.01	58.79	69.04	68.56	67.34
V	61.43	61.50	68.69	68.74	57.31
T	84.70	84.23	84.36	84.08	74.80
A+V	63.35	63.32	72.03	71.90	68.75
T+A	86.71	86.79	86.47	86.42	78.93
T+V	87.14	86.99	86.73	86.45	77.52
PaSE(T+A+V)	88.32	88.25	88.10	87.96	80.47
$\Delta(\%)$	$\uparrow 3.62$	$\uparrow 4.02$	$\uparrow 3.74$	$\uparrow 3.88$	$\uparrow 5.67$

Table 4: Performance comparison under different modality combinations across three benchmarks. Only ‘negative/positive’ results are reported for Acc-2 and F1, where T, A, and V denote Text, Audio, and Video, respectively.

essential for robust sentiment prediction. Further, removing the cross-modal alignment (CAL) module causes Acc-2 to decline by 1.52% and 1.41% on the two datasets, respectively, demonstrating its role in facilitating semantic coherence prior to fusion. Lastly, the Shapley-based gradient modulation (SGM) contributes most significantly; its removal results in 2.85% and 3.07% performance drop on MOSI and MOSEI, underscoring its ability to mitigate modality dominance and enhance collaborative optimization in the later training stages. These results collectively affirm the effectiveness and necessity of each component in PaSE.

Ablation on Fusion Strategy

To evaluate the effectiveness of our proposed Prototype-Gated Fusion (PGF) mechanism, we conduct a comparative analysis with several alternative fusion strategies, including ‘Sum’ (simple addition), ‘Con’ (MLP-based concatenation), and ‘Att’ (attention-based fusion). As shown in Table 3, the ‘Sum’ fails to capture modality-specific semantics, leading to significant information loss and degraded performance. Although ‘Con’ and ‘Att’ offer stronger representational capacity, they struggle to resolve modality competition effectively. In contrast, PGF explicitly accounts for the varying contributions of different modalities and enables more discriminative and balanced feature fusion, consistently outperforming baseline methods across multiple evaluation metrics.

Modality Fusion Gain Evaluation

To systematically assess PaSE’s ability to alleviate modality competition and enhance multimodal synergy, we evaluate its performance under different modality combinations

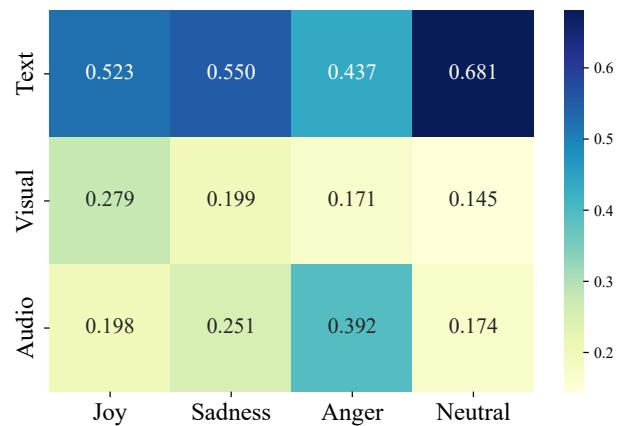


Figure 4: Heatmap of modality contributions for four emotion categories on IEMOCAP. Text, Visual, and Audio modalities show varying influence, with Text generally dominating, especially in Neutral and Sadness.

across three datasets. As shown in Table 4, incorporating additional modalities consistently improves performance, indicating PaSE’s capacity to exploit cross-modal complementarity. Specifically, as illustrated in Figure 1, fusing the visual modality (V+T) yields a 2.76% improvement—significantly surpassing GCNet’s 0.60% under the same condition. Furthermore, under the full-modality setting (T+A+V), PaSE achieves an average gain of 4.02% over all bimodal variants, effectively avoiding the diminishing returns observed in prior approaches such as MCTN (Pham et al. 2019) and MMIN (Fernando et al. 2021). Notably, the results in Table 4 also demonstrate PaSE’s **robustness under missing-modality scenarios**, where the model maintains competitive performance even with partial inputs, confirming its strong generalization and resilience to modality absence.

In-depth Analysis

To further investigate the effectiveness of PaSE, we conduct in-depth analyses by exploring the following key questions:

Q1. How does prototype-guided alignment enhance the discriminative structure of the fused feature space? We visualize the multimodal representations learned by PaSE on the CMU-MOSI test set using t-SNE, and compare them with SelfMM (Yu et al. 2021) and EUAR (Gao et al. 2024). Sentiment labels are categorized as *Positive*, *Neutral*, and *Negative*. As shown in Figure 3, the fused representations of SelfMM and EUAR exhibit substantial overlap across sentiment categories, leading to indistinct boundaries and weak intra-class cohesion. In contrast, PaSE forms compact and clearly separated clusters, indicating stronger discriminability in the fused space. This improvement stems from PaSE’s dual alignment strategy: prototype calibration enhances intra-modality discrimination by refining class-specific structures, while entropic optimal transport enforces cross-modal semantic consistency. Jointly, these mechanisms organize representations around well-defined semantic prototypes, resulting in improved class separability and robust sentiment prediction.

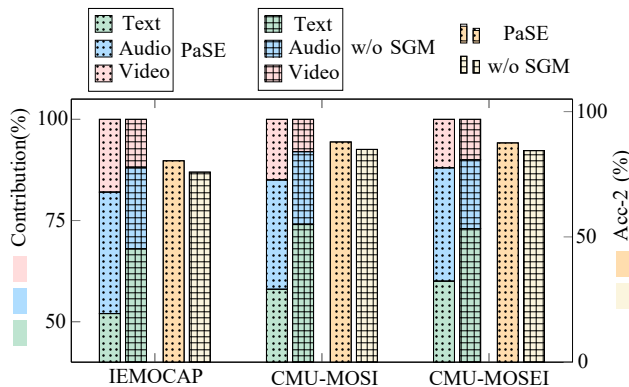


Figure 5: The influence of SGM on different modalities and the overall performance. The dual y-axis shows the average contribution of each modality (left) and the corresponding Acc-2 score (right).

Q2. How does the Prototype-Gated Fusion module influence modality fusion? We perform a fine-grained analysis of modality attention weights across different emotion categories on the IEMOCAP dataset. Specifically, we group the test samples by emotion class and compute the average attention weight assigned to each modality by the Prototype-Gated Fusion (PGF) module. As illustrated in Figure 4, the PGF module demonstrates strong context-aware fusion behavior by dynamically adjusting modality contributions based on the emotional semantics of each instance.

For example, in the case of Anger, the model assigns the highest average weight to the audio modality (0.392), effectively capturing prosodic and paralinguistic cues such as pitch, intensity, and vocal tension—hallmarks of emotionally intense speech. In contrast, for Neutral expressions, the text modality receives the dominant weight (0.681), indicating that verbal semantics play a more critical role when emotional cues are less pronounced. This context-sensitive behavior aligns well with human perceptual intuition, highlighting PGF’s ability to prioritize the most informative modality for each emotional state.

Q3. How does Shapley-guided gradient modulation alleviate modality competition? We analyzed the effect of the SGM module on three benchmark datasets. Figure 5 illustrates the comparison of modality contributions and the overall Acc-2 score. Without SGM, the text modality overwhelmingly dominates the fusion process, indicating strong modality competition and potential overfitting to a single modality. After incorporating SGM, the dominance of the text modality is significantly reduced. Contributions from the visual and audio modalities increase correspondingly, resulting in a more balanced and cooperative fusion process. Furthermore, the overall fusion accuracy improves consistently during training when SGM is applied. These results validate that SGM effectively alleviates modality competition by encouraging cooperative gradient interactions among modalities. Consequently, this leads to improved modality balance and enhances the robustness and generalization ability of the multimodal fusion model.

	Models	IEMOCAP WA(%)	MOSI Acc-2(%)	MOSEI Acc-2(%)
LLMs	Qwen2-7B	40.15	81.69	80.80
	Llama-3-8B	38.60	81.44	80.71
	Internlm-2.5-7B	39.49	84.27	82.95
MLLMs	mPLUG-Owl	55.17	72.33	72.21
	Qwen2-VL-7B	34.23	79.84	79.39
	LLaVA-OV-7B	40.09	<u>87.10</u>	<u>86.73</u>
	VideoLLaMA2-7B	38.86	79.17	78.54
	GPT-4o-mini	46.33	86.54	86.06
Bert-base	PaSE (Ours)	80.47	88.32	88.10

Table 5: Comparison of our PaSE model with several LLMs and MLLMs on IEMOCAP, MOSI, and MOSEI datasets. Metrics are Weighted Accuracy (WA) for IEMOCAP and binary accuracy (Acc-2) for MOSI and MOSEI.

Q4. Do general-purpose LLMs generalize well enough for multimodal sentiment analysis tasks, or is specialized adaptation still necessary? We conduct comparative experiments to evaluate the effectiveness of LLMs and multimodal LLMs (MLLMs) on multimodal sentiment analysis. While these models demonstrate impressive generalization across a wide range of NLP tasks, they *underperform in domain-specific scenarios*, where modality coordination and emotional nuance play a critical role. As shown in Table 5, LLM-based approaches achieve substantially lower performance than specialized multimodal architectures, particularly on the IEMOCAP dataset, which involves spontaneous dialogue, speaker variation, and subtle emotional ambiguity. These results highlight a fundamental limitation of general-purpose models: without dedicated adaptation or modality-aware optimization, they fail to capture the fine-grained inter-modal dependencies essential for robust affective reasoning. In contrast, PaSE, built upon a lightweight BERT-base backbone, consistently outperforms these general-purpose models across all benchmarks. By incorporating prototype-guided alignment and Shapley-based optimization, PaSE introduces task-specific inductive biases that strengthen cross-modal interaction, improve semantic consistency, and yield more stable and interpretable affective representations.

Conclusion

This paper proposes the PaSE framework to dynamically balance modality collaboration and competition in multimodal sentiment analysis. We utilize prototype calibration to perform intra-modality alignment, strengthening the representation of non-dominant modalities while preventing their suppression. For cross-modality alignment, an entropic optimal transport mechanism is employed to achieve effective semantic alignment between modalities. Furthermore, a dual-phase optimization strategy is designed: the first phase adaptively fuses features based on modality contributions to enhance overall representation, and the second phase applies Shapley-based gradient modulation to mitigate dominance-induced suppression. Extensive experiments on CMU-MOSI, CMU-MOSEI and IEMOCAP datasets, together with thorough ablation studies, validate the robustness and efficacy of PaSE and its core components.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62176187).

References

- Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. 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.
- Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.
- Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 960–964. IEEE.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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.
- Ding, Y.; He, K.; Li, B.; Zheng, L.; He, H.; Li, F.; Teng, C.; and Ji, D. 2025. Zero-Shot Conversational Stance Detection: Dataset and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2025*, 3221–3235.
- Fan, C.; Zhu, K.; Tao, J.; Yi, G.; Xue, J.; and Lv, Z. 2024a. Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 16(1): 207–222.
- Fan, S.; Lin, C.; Li, H.; Lin, Z.; Su, J.; Zhang, H.; Gong, Y.; Guo, J.; and Duan, N. 2022. Sentiment-Aware Word and Sentence Level Pre-training for Sentiment Analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4984–4994.
- Fan, Y.; Xu, W.; Wang, H.; Liu, J.; and Guo, S. 2024b. Detached and interactive multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5470–5478.
- Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20029–20038.
- Fernando, M.-P.; Cèsar, F.; David, N.; and José, H.-O. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7): 3217–3258.
- Fujimori, N.; Endo, R.; Kawai, Y.; and Mochizuki, T. 2020. Modality-specific learning rate control for multimodal classification. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, 412–422.
- Gao, Z.; Hu, D.; Jiang, X.; Lu, H.; Shen, H. T.; and Xu, X. 2024. Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9650–9659.
- Gul, F. 1989. Bargaining foundations of Shapley value. *Econometrica: Journal of the Econometric Society*, 81–95.
- Gushchin, N.; Kolesov, A.; Korotin, A.; Vetrov, D. P.; and Burnaev, E. 2023. Entropic neural optimal transport via diffusion processes. *Advances in Neural Information Processing Systems*, 36: 75517–75544.
- Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.
- He, K.; Ding, Y.; Li, B.; Wang, H.; Li, F.; Teng, C.; and Ji, D. 2025a. Harnessing Dimensional Contrast and Information Compensation for Sentence Embedding Enhancement. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.
- He, K.; Ding, Y.; Wang, H.; Li, F.; Teng, C.; and Ji, D. 2025b. DALR: Dual-level Alignment Learning for Multimodal Sentence Representation Learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, 3586–3601.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.
- He, X.; Liang, H.; Peng, B.; Xie, W.; Khan, M. H.; Song, S.; and Yu, Z. 2025c. MSamba: Exploring Multimodal Sentiment Analysis with State Space Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1309–1317.
- Huang, Y.; Lin, J.; Zhou, C.; Yang, H.; and Huang, L. 2022. Modality competition: What makes joint training of multimodal network fail in deep learning?(provably). In *International conference on machine learning*, 9226–9259.
- Kassraie, P.; Pooladian, A.-A.; Klein, M.; Thornton, J.; Niles-Weed, J.; and Cuturi, M. 2024. Progressive entropic optimal transport solvers. *Advances in Neural Information Processing Systems*, 37: 19561–19590.
- Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6631–6640.
- Lin, J.; Wang, Y.; Xu, Y.; and Liu, Q. 2025. Semi-IIN: Semi-supervised Intra-inter modal Interaction Learning Network for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1411–1419.
- Mai, S.; Zeng, Y.; Zheng, S.; and Hu, H. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3): 2276–2289.
- Peng, X.; Wei, Y.; Deng, A.; Wang, D.; and Hu, D. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8238–8247.

- Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6892–6899.
- Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, 439–448.
- Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, 2359.
- Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558.
- Wang, H.; He, K.; Li, B.; Chen, L.; Li, F.; Han, X.; Teng, C.; and Ji, D. 2024. Refining and Synthesis: A Simple yet Effective Data Augmentation Framework for Cross-Domain Aspect-based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, 10318–10329.
- Wang, W.; Tran, D.; and Feiszli, M. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12695–12705.
- Wu, N.; Jastrzebski, S.; Cho, K.; and Geras, K. J. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, 24043–24055.
- Xu, W.; Jiang, H.; and Liang, X. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 438–446.
- Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023. 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.
- Yang, K.; Xu, H.; and Gao, K. 2020. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 521–528.
- Yang, Y.; Dong, X.; and Qiang, Y. 2024. CLGSI: a multi-modal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2099–2110.
- Yu, J.; and Jiang, J. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.
- Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10790–10797.
- Yu, Y.; Zhao, M.; Qi, S.-a.; Sun, F.; Wang, B.; Guo, W.; Wang, X.; Yang, L.; and Niu, D. 2023. ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13610–13624.
- Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.
- Zhai, W.; Qi, H.; Zhao, Q.; Li, J.; Wang, Z.; Wang, H.; Yang, B.; and Fu, G. 2024. Chinese MentalBERT: Domain-Adaptive Pre-training on Social Media for Chinese Mental Health Text Analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, 10574–10585.
- Zhang, T.; Zhang, X.; Zhao, J.; Zhou, L.; and Jin, Q. 2024. ESCoT: Towards Interpretable Emotional Support Dialogue Systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13395–13412.
- Zhang, X.; Wei, W.; and Zou, S. 2025. Modal Feature Optimization Network with Prompt for Multimodal Sentiment Analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, 4611–4621.
- Zhou, Y.; Wang, X.; Chen, H.; Duan, X.; and Zhu, W. 2023. Intra-and inter-modal curriculum for multimodal learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3724–3735.
- Zhu, A.; Hu, M.; Wang, X.; Yang, J.; Tang, Y.; and Ren, F. 2024. KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5732–5741.
- Zhuang, Y.; Zhang, Y.; Hu, Z.; Zhang, X.; Deng, J.; and Ren, F. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1800–1809.