Beyond Static Alignment: Adaptive Arbitration for Semantic Incongruence in Semi-Supervised Multimodal Sentiment Analysis

Huicong Li* **Xiangbo Ji*** **Wei Wu†**
School of Computer Science, Inner Mongolia University
{lihuicong,jixiangbo}@mail.imu.edu.cn, cswuwei@imu.edu.cn

Abstract

Multimodal sentiment analysis is fundamentally challenged by semantic incongruence, where ambiguous visual signals often conflict with explicit textual cues. In semi-supervised scenarios, naively fusing such noisy features contaminates the joint representation, while conventional static alignment strategies fail to effectively arbitrate conflicting modalities in this task, leading to error reinforcement during self-training. To this end, we propose a novel Adaptive Arbitration for Semantic Incongruence (A2SI) framework for semi-supervised multimodal sentiment analysis, which emphasizes stable cross-modal representations and reliable supervision. Specifically, we first constrain unreliable visual representations by leveraging the reliable textual modality as an anchor to align divergent embeddings and reduce representation noise. Based on this, we further consider the reliability of supervision signals and calibrate pseudo-labels by adaptively weighting evidentiary confidence from heterogeneous views. Finally, to prevent error accumulation caused by unreliable samples, we introduce a progressive arbitration mechanism that verifies pseudo-labeled data from dual perspectives, enabling the model to dynamically balance sample diversity and label purity throughout self-training. Extensive experiments on the MVSA-Single and MVSA-Multiple datasets demonstrate that A2SI consistently outperforms state-of-the-art methods under label-limited settings.

1 Introduction

With the rapid growth of social media and digital communication platforms, users increasingly express their opinions and emotions through multimodal content, such as text and images. This trend has driven extensive research in Multimodal Sentiment Analysis (MSA) (Zhu et al., 2025; Liu et al., 2024; Wang et al., 2025), which aims to infer user sentiment by integrating heterogeneous cues into a unified representation. Benefiting from cross-modal complementarity, MSA has been widely explored in areas such as social media analysis, user profiling, and human-computer interaction (Yang et al., 2023; Wu et al., 2024). However, given the scarcity of high-quality annotations in practice, extending MSA to semi-supervised scenarios is desirable but remains hindered by the inherent unevenness of modality quality and the complexity of pseudo-label generation.

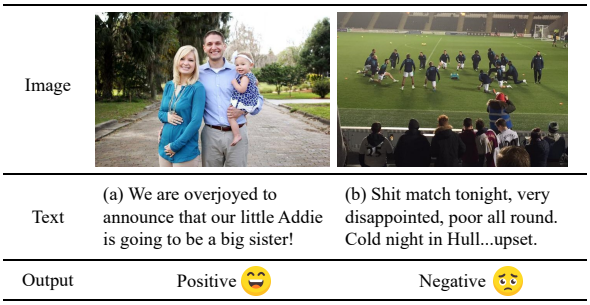


Figure 1: Examples of multimodal sentiment samples illustrating semantic alignment and incongruence between image and text modalities.

Visual signals can perfectly align with textual sentiment in ideal cases as shown in Figure 1(a), but in scenarios characterized by visual ambiguity, textual sentiment is often explicit, whereas visual cues tend to be high-entropy and context-dependent (Huang et al., 2026; Zhao et al., 2025; Wu et al., 2026). As illustrated in Figure 1(b), the text provides a clear semantic anchor, while the visual view is dominated by task-irrelevant background patterns. Without explicit semantic regularization, visual encoders are prone to overfitting such dominant but sentiment-agnostic cues, causing the visual embedding to diverge from the underlying sentiment manifold . When these unstable visual representations are directly fused with reliable textual features, they dilute the discriminative signal and degrade the robustness of the joint rep-

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41021–41032
July 2-7, 2026

resentation.

In semi-supervised learning (SSL), existing approaches often derive supervision signals primarily from fused representations (Yang et al., 2019, 2021b), implicitly assuming that cross-modal fusion yields the most reliable predictions. However, this assumption can break down under semantic incongruence, where a noisy view can corrupt the fused representation and lead to incorrect joint predictions, even when unimodal evidence remains reliable. More specifically, such incongruence may induce a progressive degradation process, where corrupted predictions result in noisy pseudo-labels(Feng et al., 2022; Paul et al., 2019). As these labels are iteratively reused during self-training, they can further contribute to error accumulation over time(Chen et al., 2022a). To mitigate such noise, prior methods typically rely on static confidence thresholds (Sohn et al., 2020). These rigid criteria lack the flexibility to arbitrate conflicting heterogeneous views (Zhang et al., 2021; Xu et al., 2021). As a result, they tend to discard informative hard samples while failing to suppress high-confidence errors induced by modality bias, making it difficult to balance sample diversity with label purity during self-training.

To overcome the limitations of static alignment under semantic incongruence, we propose Adaptive Arbitration for Semantic Incongruence (A2SI), a semi-supervised framework for image-text sentiment analysis, which consists of three components: Fusion-Guided Modality Regularization (FGMR), Reliability-Aware Calibration (RAC), and Dual-View Arbitration Distillation (DVAD). Specifically, for each text-image pair, we first utilize FGMR to mitigate visual noise by leveraging the reliable textual semantics as a semantic anchor, thereby imposing explicit semantic constraints on visual encoders and stabilizing modality-specific representations. Building on these stabilized representations, the RAC module adaptively integrates unimodal and multimodal evidence by re-weighting their confidence, generating calibrated pseudo-label supervision. Finally, we introduce DVAD, which acts as a progressive arbitration mechanism that verifies pseudo-labeled samples from dual perspectives, enabling the model to effectively arbitrate conflicting cues and dynamically balance sample diversity and label purity during self-training.

The contributions can be summarized as follows:

- We propose a novel semi-supervised framework, Adaptive Arbitration for Semantic Incongruence (A2SI), for multimodal sentiment analysis, which systematically addresses semantic incongruence by improving representation stability and pseudo-label reliability in self-training.
- We design an integrated learning strategy that stabilizes noisy cross-modal representations through text-guided semantic constraints and calibrates pseudo-label supervision by modeling the reliability of heterogeneous evidence, mitigating both representation noise and supervision bias.
- We introduce a progressive arbitration mechanism that adaptively verifies training samples from complementary perspectives, effectively preventing error accumulation and enabling a dynamic balance between sample diversity and label purity during self-training.
- Extensive experiments on two benchmark datasets, including MVSA-Single and MVSA-Multiple, show that our approach outperforms the state-of-the-art methods across various label-limited settings.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) integrates heterogeneous signals via feature-level fusion, evolving from early tensor-based and attention mechanisms (Zadeh et al., 2017; Xu and Mao, 2017) to recent frameworks that address weak correlations (Li et al., 2025) or leverage vision-language model priors (Huang et al., 2024a; Ling et al., 2022) for enhanced alignment. Complementing this, decision-level fusion aggregates unimodal predictions, where recent approaches like tag-assisted mechanisms (Zeng et al., 2023) employ importance weighting to robustly handle incomplete data patterns. However, existing methods mainly focus on fusion mechanisms or missing modality recovery, which is inadequate to prevent the semantic divergence of visual representations caused by task-irrelevant background noise.

2.2 Semi-supervised Multimodal Learning

Semi-supervised multimodal learning leverages unlabeled data via self-training and consistency regularization (Zhang et al., 2016; Sirbu et al., 2022),

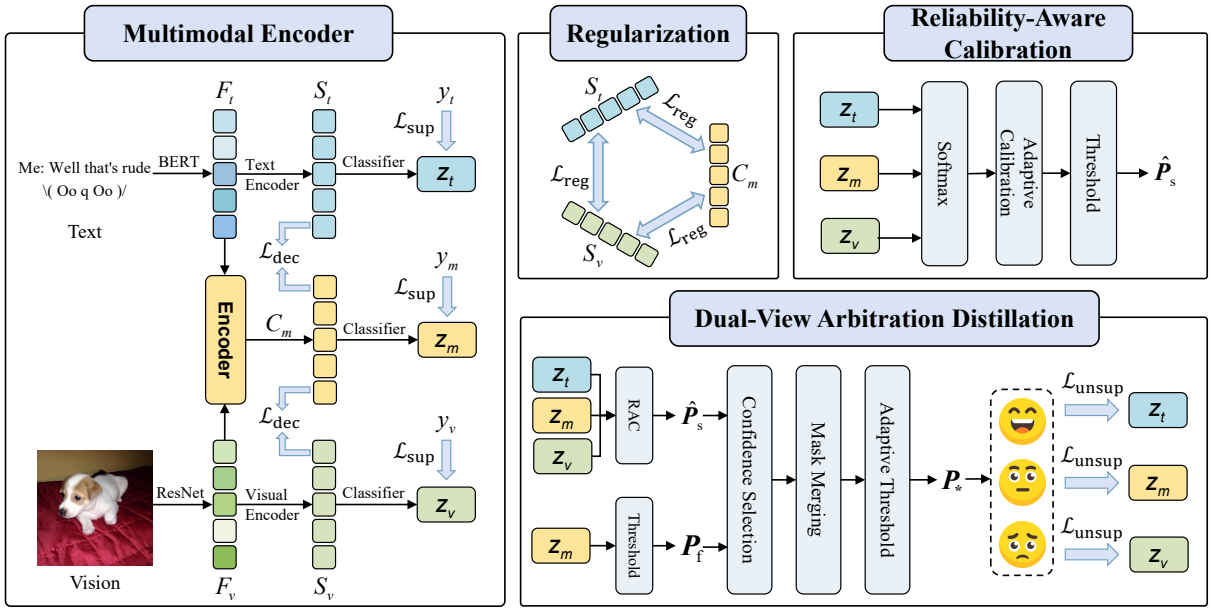


Figure 2: Overview of the proposed Adaptive Arbitration for Semantic Incongruence (A2SI) framework. Visual and textual inputs are encoded into modality-specific and fused representations, followed by Fusion-Guided Modality Regularization (FGMR) and Classification. Unlabeled samples are refined through the Reliability-Aware Calibration (RAC) and the Dual-View Arbitration Distillation (DVAD).

often enforcing cross-modal alignment through co-training paradigms (Liang et al., 2020). To mitigate unreliable predictions, recent works focus on refining supervision quality, evolving from class-balanced selection strategies (Chen et al., 2023) to dynamic learnable filter networks (Yuan et al., 2024) that effectively distill knowledge by filtering noisy pseudo-labels. Considering the reliability asymmetry across different views, our method recalibrates prediction confidence to adaptively arbitrate between conflicting predictions, thereby correcting biased pseudo-labels without relying on rigid thresholds.

3 Method

3.1 Overall Architecture

We present an overview of our framework in Figure 2, which consists of four core modules. (1) The Multimodal Feature Encoder maps visual and text inputs into disentangled modality-specific and fusion-oriented semantic spaces. (2) The Fusion-Guided Modality Regularization (FGMR) module stabilizes training by anchoring unimodal representations to the reliable fusion center for semantic coherence. (3) The Reliability-Aware Calibration (RAC) module recalibrates pseudo-labels by synthesizing evidence from all branches to filter noisy modalities. (4) The Dual-View Arbitration Distillation (DVAD) module adaptively arbitrates between the calibrated and fusion views to distill high-quality supervision signals. Next, we describe these modules in detail.

3.2 Multimodal Feature Encoding

Our framework processes image-text pairs from both a labeled batch $\mathcal{X} = \{(x_{b,v}, x_{b,t}, y_b)\}_{b=1}^{B}$ and an unlabeled batch $\mathcal{U} = \{(u_{b,v}, u_{b,t})\}_{b=1}^{\mu B}$. Visual and text inputs are first processed by distinct backbones to obtain raw representations F_v and F_t. These are then mapped into two sets of features: modality-specific features S_v, S_t via individual encoders, and fusion-oriented features C_v, C_t. The latter are summed to form the unified multimodal representation C_m:

$$\begin{aligned} S_v = f_v(F_v), \quad S_t = f_t(F_t), \\ C_m = f_m(F_v) + f_m(F_t). \end{aligned} \tag{1}$$

Classification & Supervised Loss. With the semantic space aligned, the stabilized features S_v, S_t, and C_m are fed into their respective classifiers, yielding predicted logits $z_{b,v}$, $z_{b,t}$ and $z_{b,m}$ for each sample b. For labeled data in $\mathcal{X}$, utilizing the ground-truth label y_b, the model is optimized via a multi-branch supervised loss based on the

standard cross-entropy function ℓ_{CE}:

$$\mathcal{L}_{\text{sup}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in\{v,t,m\}} \ell_{\text{CE}}\left(z_{b,k},\, y_b\right). \quad (2)$$

3.3 Fusion-Guided Modality Regularization

As aforementioned, visual modalities in real-world scenarios are often plagued by ambiguity and high entropy. Lacking explicit semantic constraints, the visual encoder is prone to overfitting task-irrelevant background patterns, causing the visual embedding S_v to diverge from the true sentiment manifold. To rectify this divergence, we propose Fusion-Guided Modality Regularization (FGMR).

FGMR serves as a lightweight constraint mechanism that stabilizes modality-specific representations. Specifically, it encourages semantic coherence among the embeddings by constraining their pairwise distances:

$$\begin{aligned}\mathcal{L}_{\text{reg}} = &\|S_t - S_v\|_2 \\ &+ \alpha\left(\|S_t - C_m\|_2 + \|S_v - C_m\|_2\right).\end{aligned} \quad (3)$$

Crucially, the fusion representation C_m explicitly integrates the robust semantics from the textual branch, thereby functioning as a stable semantic anchor. By minimizing $\|S_v - C_m\|_2$, the framework acts as a semantic rectifier: it compels the visual encoder to suppress irrelevant background noise and re-align its feature distribution with the robust sentiment context, effectively mitigating the feature divergence problem.

Feature Decoupling. To further disentangle multimodal and modality-specific representations, we incorporate an auxiliary feature-decoupling loss (Li et al., 2023). This component complements FGMR by preventing shortcut correlations, ensuring that sentiment-relevant fused features remain discriminative while preserving modality-unique information. The decoupling objective is thus composed of reconstruction, cycle-consistency, margin-based, and orthogonality terms:

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}} + \gamma(\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{ort}}). \quad (4)$$

3.4 Reliability-Aware Calibration Module

As analyzed previously, pseudo-labels derived solely from a single fusion classifier are vulnerable under asymmetric modality contributions. In such cases, a noisy modality can contaminate the joint representation, causing the fusion prediction to deviate from the truth even when valid evidentiary support is preserved in unimodal branches. To mitigate this contamination and fully leverage unimodal expertise, we introduce the Reliability-Aware Calibration (RAC) module. This module recalibrates predictions by synthesizing evidence from visual, textual, and fusion logits z_v, z_t, and z_m.

Reliability Quantification. To quantify the evidentiary quality of each branch $i \in \{v, t, m\}$, we first apply temperature scaling to enhance the discriminability of class probabilities and extract the maximum prediction confidence:

$$i_{\max} = \max\left(\text{Softmax}\left(\frac{z_i}{T}\right)\right), \quad (5)$$

where T controls the sharpness of the distribution, and $i_{\max}$ serves as a dynamic indicator of how trustworthy the i-th modality is for the current instance.

Adaptive Calibration & Filtering. These confidence cues are subsequently used to compute adaptive modality weights, yielding the corrected pseudo-label $\tilde{P}_s$:

$$\tilde{P}_s = \sum_{i\in\{v,t,m\}} \frac{i_{\max}}{v_{\max} + t_{\max} + m_{\max}} P_i. \quad (6)$$

This formulation naturally shifts the decision focus to the most reliable branch. In scenarios where fusion is compromised by noise, for example background-heavy images, RAC allows clean unimodal cues, for example explicit text, to override the contaminated fusion output. Finally, a soft confidence filter is applied:

$$\hat{\boldsymbol{P}}_s = \begin{cases} \widetilde{P}_s, & \max(\widetilde{P}_s) \geq \tau_{\text{soft}}, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

which yields the corrected pseudo-label candidate $\hat{\boldsymbol{P}}_s$ for the subsequent adaptive arbitration stage.

3.5 Dual-View Arbitration Distillation

Traditional approaches often rely on a single static gate to select pseudo-labels, overlooking the dynamic reliability shifts between modalities. To address this, the Dual-View Arbitration Distillation (DVAD) module is designed to adaptively distill high-quality supervision signals by arbitrating between two complementary experts: the calibration-based pseudo-label $\hat{\boldsymbol{P}}_s$ generated by the RAC module, and the fusion-based prediction $\boldsymbol{P}_f$ derived from the fusion logits z_m via Softmax.

Adaptive Arbitration Strategy. For each unlabeled sample, the two views yield predictions

$\hat{\boldsymbol{P}}_s = (p_s, c_s, m_s)$ and $\boldsymbol{P}_f = (p_f, c_f, m_f)$, where p denotes class probabilities, c represents confidence scores, and $m \in \{0, 1\}$ is the binary selection mask. DVAD evaluates their relative reliability on a per-sample basis. The final pseudo-label target p_{sel} is selected based on a confidence margin δ, which allows flexible arbitration between the calibrated and fusion views:

$$p_{\text{sel}} = \begin{cases} \arg\max p_s, & c_f - c_s < \delta_t, \\ \arg\max p_f, & \text{otherwise.} \end{cases} \tag{8}$$

To stabilize training, δ_t follows a linear warm-up schedule over epochs t:

$$\delta_t = \delta_{base} + (\delta_{max} - \delta_{base}) \times \min\left(1, \frac{t}{T_{warmup}}\right). \tag{9}$$

This curriculum-style schedule gradually increases the margin as the model converges. Initially, a smaller margin allows the fusion branch to contribute more freely. By imposing a stricter constraint in later stages, it defers to the calibrated view, ensuring the fusion view is selected only when a significant confidence gap exists ($c_f \geq c_s + \delta_t$), thus preventing instability during self-training.

Progressive Tightening & Distillation. To ensure the framework transitions from robust exploration to high-precision exploitation, we apply a Progressive Tightening Filter. An unlabeled sample is accepted only if the average confidence of both views exceeds a adaptive threshold $\gamma(t)$:

$$m_{\text{final}} = \mathbb{I}\left(\frac{m_s c_s + m_f c_f}{2} \geq \gamma(t)\right), \tag{10}$$

where the threshold $\gamma(t)$ increases linearly from $\gamma_{\min}$ to $\gamma_{\max}$ as the epoch t advances. This design enforces a strict purity constraint to eliminate low-confidence noise as the model converges.

Finally, the filtered probabilities $\boldsymbol{P}_* = \{p_{\text{sel}}\}$ are obtained and incorporated into the training process. For an unlabeled batch of size μB, these targets form the supervision set $\mathcal{T}_u = \{(p_b^{\text{sel}}, m_{b,\text{final}})\}_{b=1}^{\mu B}$. This set guides the unsupervised learning by aligning the unlabeled logits $z_{b,k}^{\text{unsup}}$ across all branches with the selected pseudo-labels. Implemented via the unsupervised loss, this process serves as a self-distillation mechanism to transfer reliable knowledge across all modalities $k \in \{v, t, m\}$:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} m_{b,\text{final}} \sum_{k} \ell_{\text{CE}}\left(z_{b,k}^{\text{unsup}}, p_b^{\text{sel}}\right). \tag{11}$$

Finally, the overall training objective integrates the supervised learning, unsupervised self-distillation, feature decoupling, and regularization constraints. With λ_{reg} controlling the relative strength of the stabilization constraint, the total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}} + \mathcal{L}_{\text{dec}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{12}$$

4 Experiments

4.1 Dataset Dataset and Evaluation Metrics

Datasets. The original MVSA dataset (Niu et al., 2016) includes two versions with sentiment polarity labels for both image and text modalities: MVSA-Single, which contains 4,869 data pairs, and MVSA-Multiple, which contains 19,600 data pairs. These sentiment annotations are provided independently for the textual and visual modalities, leading to a notable degree of cross-modal inconsistency. To address this issue, we evaluate our method on the relabeled MVSA-Single and MVSA-Multiple datasets released by (Xia et al., 2025), which introduce independent annotations for image, text, and overall multimodal sentiment, thereby providing a more coherent and reliable supervision signal for multimodal learning.

Evaluation metrics. We adopt accuracy as the primary evaluation metric. Additionally, to monitor the reliability of the self-training process, we report Pseudo-Label Accuracy (PLA), calculated against the ground truth of unlabeled data, and Utilization Ratio (UR), defined as the proportion of unlabeled samples selected for training. Following previous semi-supervised works (Sun et al., 2020; Yang et al., 2019; Sirbu et al., 2022; Xia et al., 2025), we report the mean and variance of the 5-fold experiment accuracy.

4.2 Implementation Details.

All experiments are implemented in PyTorch 2.5.0 and conducted on an NVIDIA GTX 3090 GPU. We adopt ResNet-18 as the visual encoder and BERT-base as the text encoder. For the RAC module, the temperature scaling parameter T is set to 0.7. For the training-scheduled margin δ_t in the DVAD module, we set δ_{base} to 0.02, δ_{max} to 0.05,

Model	Venues	MVSA-Single			MVSA-Multiple		
		50	200	600	100	600	1500
MGNNS	ACL 2021	$43.62_{\pm 0.68}$	$52.86_{\pm 0.74}$	$60.45_{\pm 0.64}$	$46.77_{\pm 0.71}$	$53.60_{\pm 0.58}$	$59.12_{\pm 0.76}$
Hgraph-CL	COLING 2022	$41.34_{\pm 0.74}$	$45.69_{\pm 0.54}$	$51.47_{\pm 0.47}$	$43.04_{\pm 0.43}$	$48.37_{\pm 0.65}$	$54.26_{\pm 0.69}$
INIT	TMM 2022	$44.14_{\pm 0.87}$	$50.48_{\pm 0.76}$	$58.37_{\pm 0.64}$	$45.65_{\pm 1.29}$	$52.60_{\pm 0.99}$	$57.33_{\pm 0.73}$
CLMLF	NAACL 2022	$53.55_{\pm 0.67}$	$57.02_{\pm 0.52}$	$62.33_{\pm 0.33}$	$48.27_{\pm 0.87}$	$54.12_{\pm 0.82}$	$57.76_{\pm 1.20}$
DPRN	TMM 2023	$46.52_{\pm 0.72}$	$53.64_{\pm 0.62}$	$58.02_{\pm 0.84}$	$46.20_{\pm 0.52}$	$53.67_{\pm 0.48}$	$56.97_{\pm 0.68}$
MVCN	ACL 2023	$42.60_{\pm 0.94}$	$52.52_{\pm 0.85}$	$59.12_{\pm 0.74}$	$50.12_{\pm 0.67}$	$57.59_{\pm 0.94}$	$60.33_{\pm 1.03}$
DMD	CVPR 2023	$46.24_{\pm 0.65}$	$55.69_{\pm 0.75}$	$61.67_{\pm 0.57}$	$52.48_{\pm 0.72}$	$57.48_{\pm 0.59}$	$63.77_{\pm 0.67}$
CMGCN	COLING 2024	$45.82_{\pm 0.45}$	$53.84_{\pm 0.58}$	$60.24_{\pm 0.42}$	$49.56_{\pm 0.47}$	$56.25_{\pm 0.62}$	$59.24_{\pm 0.75}$
SPFVTE	ICMR 2024	$53.09_{\pm 0.63}$	$57.87_{\pm 0.76}$	$64.60_{\pm 0.50}$	$60.79_{\pm 0.84}$	$64.06_{\pm 0.79}$	$66.35_{\pm 0.56}$
ArkMSA	ICME 2024	$51.66_{\pm 0.69}$	$62.31_{\pm 0.52}$	$66.55_{\pm 0.76}$	$61.52_{\pm 0.53}$	$65.17_{\pm 0.83}$	$66.27_{\pm 0.49}$
MixText	ACL 2020	$58.16_{\pm 0.68}$	$61.48_{\pm 0.62}$	$63.14_{\pm 0.62}$	$58.47_{\pm 0.48}$	$61.39_{\pm 0.67}$	$65.01_{\pm 0.66}$
SAT	EMNLP 2022	$57.09_{\pm 0.57}$	$60.33_{\pm 0.63}$	$62.87_{\pm 0.45}$	$60.70_{\pm 0.64}$	$63.47_{\pm 0.49}$	$65.53_{\pm 0.34}$
S2-VER	ECCV 2022	$57.49_{\pm 0.78}$	$60.57_{\pm 0.69}$	$63.03_{\pm 0.74}$	$60.64_{\pm 0.45}$	$63.86_{\pm 0.59}$	$64.68_{\pm 0.36}$
CHMatch	CVPR 2023	$56.18_{\pm 0.98}$	$60.14_{\pm 0.77}$	$63.84_{\pm 0.55}$	$58.84_{\pm 0.69}$	$64.01_{\pm 0.45}$	$65.88_{\pm 0.43}$
CMML	IJCAI 2019	$52.16_{\pm 0.66}$	$55.79_{\pm 0.56}$	$61.31_{\pm 0.84}$	$58.88_{\pm 0.61}$	$60.48_{\pm 0.74}$	$63.44_{\pm 0.42}$
TCGM	ECCV 2020	$55.10_{\pm 0.68}$	$59.77_{\pm 0.47}$	$63.54_{\pm 0.44}$	$59.12_{\pm 0.77}$	$62.15_{\pm 0.53}$	$65.03_{\pm 0.41}$
FixMatchLS	COLING 2022	$60.16_{\pm 0.73}$	$62.01_{\pm 0.73}$	$63.59_{\pm 0.56}$	$62.42_{\pm 0.52}$	$64.80_{\pm 0.58}$	$66.28_{\pm 0.38}$
SCRD	CVPR 2025	$60.51_{\pm 0.32}$	$64.27_{\pm 0.72}$	$66.60_{\pm 0.59}$	$64.22_{\pm 0.47}$	$65.29_{\pm 0.36}$	$67.76_{\pm 0.39}$
A2SI	Ours	$\mathbf{61.77_{\pm 0.53}}$	$\mathbf{64.64_{\pm 0.78}}$	$\mathbf{68.75_{\pm 0.43}}$	$\mathbf{66.09_{\pm 0.39}}$	$\mathbf{67.40_{\pm 0.41}}$	$\mathbf{68.38_{\pm 0.35}}$

Table 1: The accuracy (%) and variance (%) of 5-fold cross-validation on the MVSA-Single and MVSA-Multiple datasets, comparing our Adaptive Arbitration for Semantic Incongruence (A2SI) with state-of-the-art supervised and semi-supervised approaches, where labels represent the number of annotated samples used in training.

and the warm-up period T_{warmup} to 100 epochs. For pseudo-label filtering, the confidence threshold $\gamma(t)$ linearly ramps up from an initial γ_{min} to 0.95. Specifically, we set γ_{min} to 0.15 with a 70-epoch warm-up for MVSA-Single, and 0.20 with an 80-epoch warm-up for MVSA-Multiple. For loss design, we set the FGMR regularization weight λ_{reg} to 0.05. We train the model on both datasets for 150 epochs under a semi-supervised setting, using an initial learning rate of 1×10^{-4} with cosine decay.

In our semi-supervised setting, a subset of instances from dataset is randomly selected as labeled data, while the remaining samples are treated as unlabeled by discarding their ground-truth annotations during training. The model is trained using ground-truth supervision on the labeled subset, while pseudo-labels are generated for the unlabeled data to enable semi-supervised learning. The specific numbers of labeled samples used for training are indicated in Table 1.

4.3 Comparison with State-of-the-art Methods

To comprehensively evaluate the effectiveness of our approach, we compare it with a broad range of state-of-the-art (SOTA) methods, including:

Fully supervised methods: MGNNS(Yang et al., 2021a), Hgraph-CL(Lin et al., 2022), INIT(Zhu et al., 2023), CLMLF(Li et al., 2022), DPRN(Wang et al., 2024), MVCN(Wei et al., 2023), DMD(Li et al., 2023), CMGCN(Zhang et al., 2024), SPFVTE (Huang et al., 2024b), ArkMSA(Pang et al., 2024).

Extensions of unimodal methods: MixText(Chen et al., 2020), SAT(Chen et al., 2022b), S2-VER(Jia and Yang, 2022) and CHMatch(Wu et al., 2023).

Multimodal methods: TCGM(Sun et al., 2020), CMML(Yang et al., 2019), FixMatchLS(Sirbu et al., 2022) and SCRD(Xia et al., 2025).

As shown in Table 1, our method consistently outperforms SOTA methods across all settings on both datasets. While the previous state-of-the-art method SCRD (Xia et al., 2025) demonstrates competitive performance, it relies on static thresholds to select pseudo-labels primarily from the fusion prediction. This rigid strategy limits flexibility, often discarding high-quality predictive cues preserved in unimodal branches when the fused view is compromised. In contrast, A2SI utilizes adaptive confidence weighting to dynamically synthesize evidence from all modalities and introduces a progressive arbitration mechanism to adjudicate between the calibrated and fusion results. This

Models	MVSA-S	MVSA-M
A2SI	**68.58**	**67.73**
w/o FGMR	67.96	66.92
w/o RAC	67.14	66.76
w/o DVAD	65.23	66.52

Table 2: Ablation study in terms of accuracy on MVSA-Single and MVSA-Multiple with different components.

Visual Extractor	MVSA-S	MVSA-M
ResNet-18	**68.78**	**67.68**
ResNet-50	67.76	66.30
ResNet-101	64.88	65.79
ViT-B/16	62.42	67.58
ViT-L/32	65.09	65.91
ViT-H/14	63.24	67.17

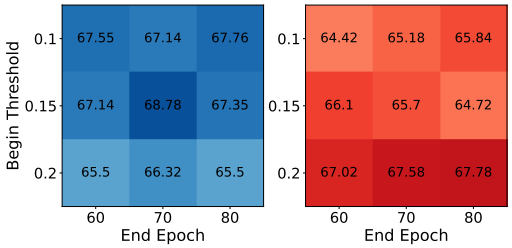
Table 3: Comparison of different visual extractor on MVSA-Single and MVSA-Multiple in terms of accuracy.

design allows for the retrieval of more reliable predictions that static gates might miss, yielding pseudo-labels of significantly higher quality. Consequently, under the 600-label setting, our approach achieves notable improvements of 2.15% on MVSA-Single and 2.11% on MVSA-Multiple compared to SOTA. Robust improvements persist even in low-data regimes, exemplified by a 1.26% gain with 50 samples on MVSA-Single and a 1.87% gain with 100 samples on MVSA-Multiple, and extend to higher-resource settings like the 1500-sample configuration, demonstrating the scalability of our design.

4.4 Ablation Studies

To ensure a fair and consistent comparison, all experiments in this section are performed using 600 labeled samples on MVSA-Single and MVSA-Multiple datasets.

Component Effectiveness. Table 2 shows that without the various components of the A2SI model, the overall accuracy decreases. Further analysis reveals that the performance is most significantly affected without DVAD, with the accuracy on MVSA-Single decreasing by 3.35%. This indicates that the adaptive arbitration between conflicting views is crucial for robust learning in label-scarce settings. As complementary modules, RAC and FGMR also play significant roles in the framework. The performance drop without RAC confirms the importance of reliability-aware calibration, while the decrease



(a) (b)

Figure 3: (a) Effect of different threshold ranges and training epochs on MVSA-Single in terms of accuracy.(b) Effect of different threshold ranges and training epochs on MVSA-Multiple in terms of accuracy.

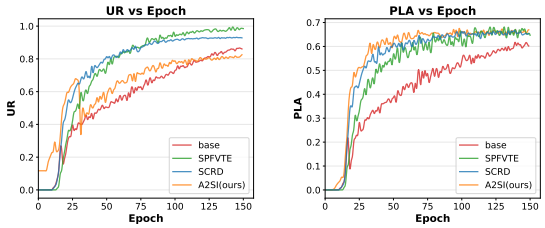
without FGMR suggests that stabilizing visual representations via fusion guidance is beneficial for enhancing model alignment.

Visual Encoder Architecture. Table 3 investigates the performance of various visual backbones, ranging from CNN-based architectures such as ResNet variants (He et al., 2016) to Vision Transformer models (Dosovitskiy, 2020). Counterintuitively, the lightest model, ResNet-18, consistently achieves the highest accuracy on both MVSA-Single and MVSA-Multiple. We observe a clear trend that increasing model complexity does not translate into performance gains; for example, scaling from ResNet-18 to ResNet-101 leads to a notable accuracy drop. These results suggest that in label-scarce semi-supervised settings, scaling up model capacity does not necessarily improve performance. Although ViT-B/16 achieves competitive results on the more complex MVSA-Multiple dataset, the advantages of larger Transformer-based backbones are difficult to fully realize under substantial irrelevant visual noise. Consequently, we adopt ResNet-18 as the default visual extractor, offering a superior balance between robustness, data efficiency, and computational cost.

Hyperparameter Sensitivity. Figure 3 examines different annealing schedules for the confidence threshold $\gamma(t)$. On the Single dataset Figure 3(a), the best performance occurs when $\gamma(t)$ increases from 0.15 to 0.95 within 70 epochs, while shorter or longer warm-up schedules degrade accuracy. For the Multiple dataset Figure 3(b), a stricter initial value 0.20 and a longer warm-up 80 epochs yield the highest accuracy, indicating greater sensitivity to noisy pseudo-labels. Overall, Single benefits from a more relaxed early threshold, whereas Multiple requires conservative early filtering.

λ_{reg}	0.01	0.03	0.05	0.07	0.09
MVSA-S	66.52	67.14	**68.78**	68.19	67.55
MVSA-M	66.32	65.70	**67.68**	66.61	65.91

Table 4: Performance comparison on MVSA-Single and MVSA-Multiple with different λ_{reg} values in terms of accuracy.



(a) (b)

Figure 4: Self-training dynamics on MVSA-Single Evolution of pseudo-label quality and quantity over epochs across different models.

Table 4 examines the sensitivity of the model to the regularization weight λ_{reg}. The results show that setting λ_{reg} to 0.05 consistently leads to the best performance on both datasets. Smaller values fail to impose sufficient semantic regularization across modalities, while larger values tend to over-regularize the model and suppress modality-specific information, resulting in degraded performance.

Quality vs. Quantity. Figure 4a reveals a clear divergence in utilization behavior between different methods. Competing approaches rapidly increase the Utilization Ratio (UR) to nearly 100%, which maximizes data usage but also leads to the inclusion of a large number of ambiguous samples as training signals. In contrast, A2SI adopts a more conservative utilization strategy, reflecting a deliberate design choice. This behavior arises from the selective filtering mechanisms of the RAC and DVAD modules, which effectively prune samples suffering from severe cross-modal ambiguity. As a result, A2SI consistently achieves higher and more stable Pseudo-Label Accuracy (PLA) throughout training, as shown in Fig. 4b. These results indicate that prioritizing pseudo-label quality over aggressive data utilization is critical for robust semi-supervised multimodal learning.

Quantitative Profiling. To validate the necessity of our dynamic arbitration strategy, we investigate the underlying characteristics of the discarded samples. We profile the retained and discarded subsets on the MVSA-Single dataset under the 600-label setting. As shown in Table 5, the discarded subset exhibits severe semantic ambiguity and unreliability. The gap is defined as the difference between the discarded and retained values. Specifically, the Pseudo-Label Accuracy (PLA) of the discarded samples drops by over 30% compared to the retained subset. In the context of semi-supervised learning, incorporating such low-quality pseudo-labels would inevitably trigger error amplification, severely contaminating the self-training process. Furthermore, the discarded samples exhibit substantially higher prediction entropy, in stark contrast to the confident predictions of the retained set.

Metric	Retained	Discarded	Gap
PLA (↑)	69.18	38.59	-30.59
Entropy (↓)	0.013	0.579	0.566

Table 5: Comparison of retained and discarded samples on MVSA-Single in terms of PLA (%) and entropy.

4.5 Visualization

As shown in Figure 5, our A2SI model demonstrates strong robustness in scenarios where visual signals are either high-entropy or semantically conflicting with the text. While visual cues in real-world data are often dominated by task-irrelevant background patterns or misleading expressions, such as the smiling face in the "negative" context of Figure 5c, A2SI correctly predicts the sentiment across these diverse cases. This confirms that our semantic regularization mechanism effectively suppresses sentiment-agnostic visual noise, forcing the visual embedding to align with the reliable textual anchor rather than overfitting to superficial visual features. By doing so, the model preserves the semantic integrity of the joint representation, particularly in scenarios where the visual modality is deceptive.

5 Conclusion

In this work, we introduced Adaptive Arbitration for Semantic Incongruence (A2SI), a semi-supervised framework that improves multimodal sentiment analysis by correcting noisy pseudo-labels and stabilizing cross-modal representations. Driven by a unified strategy that stabilizes noisy cross-modal representations, calibrates pseudo-label supervision according to evidential reliability, and adaptively resolves modality conflicts during

Figure 5: Qualitative results on the MVSA-Single dataset. The bottom labels indicate the sentiment predicted by A2SI. The model's robustness in challenging scenarios: despite high-entropy backgrounds or conflicting visual cues, A2SI correctly predicts the sentiment. This validates that our semantic regularization effectively anchors visual features to the true sentiment manifold, preventing overfitting to task-irrelevant visual patterns.

self-training, our method effectively addresses semantic incongruence by emphasizing label quality over aggressive unlabeled data utilization. Experiments on MVSA-Single and MVSA-Multiple show that A2SI achieves consistent and notable improvements over state-of-the-art methods.

Limitations

In order to resolve semantic incongruence and guarantee label purity, A2SI employs a rigorous filtering mechanism through the RAC and DVAD modules. Consequently, our method exhibits a lower Utilization Ratio (UR) compared to state-of-the-art baselines, as it aggressively filters out unlabeled instances with high cross-modal ambiguity. Although this conservative approach achieves superior Pseudo-Label Accuracy (PLA), it also implies that the model may under-utilize the valid semantic signals latent in the discarded data. In the future, we will explore more robust noise-tolerant mechanisms to mine effective supervision signals from these discarded ambiguous samples, aiming to boost data efficiency without compromising label reliability. Furthermore, while the proposed arbitration mechanism may have potential applicability to other multimodal scenarios with conflicting supervision, its generalization beyond multimodal sentiment analysis is not explored in this work.

Acknowledgements

This work is supported by the Inner Mongolia Natural Science Foundation Project No.2024MS06007 and the Inner Mongolia Science and Technology Project No.2021GG0166.

References

Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. 2022a. Debiased self-training for semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 32424–32437. Curran Associates, Inc.

Haifeng Chen, Chujia Guo, Yan Li, Peng Zhang, and Dongmei Jiang. 2023. Semi-supervised multimodal emotion recognition with class-balanced pseudo-labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pages 9556—-9560, New York, NY, USA. Association for Computing Machinery.

Hui Chen, Wei Han, and Soujanya Poria. 2022b. SAT: Improving semi-supervised text classification with simple instance-adaptive self-training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6141–6146, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130:108777.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huiting Huang, Tieliang Gong, Kai He, Jialun Wu, Erik Cambria, and Mengling Feng. 2026. Robust multimodal sentiment analysis via double information bottleneck. *Information Fusion*, 129:103964.

Qi Huang, Pingting Cai, Tanyue Nie, and Jinshan Zeng. 2024a. Clip-msa: Incorporating inter-modal dynamics and common knowledge to multimodal sentiment analysis with clip. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8145–8149. IEEE.

Shizhou Huang, Bo Xu, Changqun Li, Jiabo Ye, and Xin Lin. 2024b. A sentimental prompt framework with visual text encoder for multimodal sentiment analysis. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 638–646, New York, NY, USA. Association for Computing Machinery.

Guoli Jia and Jufeng Yang. 2022. S2-ver: Semi-supervised visual emotion recognition. In *Computer Vision – ECCV 2022*, pages 493–509, Cham. Springer Nature Switzerland.

Yangmin Li, Ruiqi Zhu, and Wengen Li. 2025. CorMulT: A Semi-Supervised Modality Correlation-Aware Multimodal Transformer for Sentiment Analysis . *IEEE Transactions on Affective Computing*, 16(03):2321–2333.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6631–6640.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, Seattle, United States. Association for Computational Linguistics.

Jingjun Liang, Ruichen Li, and Qin Jin. 2020. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 2852—-2861, New York, NY, USA. Association for Computing Machinery.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7124–7135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159, Dublin, Ireland. Association for Computational Linguistics.

Zijun Liu, Li Cai, Wenjie Yang, and Junhui Liu. 2024. Sentiment analysis based on text information enhancement and multimodal feature fusion. *Pattern Recognition*, 156:110847.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, pages 15–27, Cham. Springer International Publishing.

Ning Pang, Wansen Wu, Yue Hu, Kai Xu, Quanjun Yin, and Long Qin. 2024. Enhancing multimodal sentiment analysis via learning from large language model. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Debjit Paul, Mittul Singh, Michael A. Hedderich, and Dietrich Klakow. 2019. Handling noisy labels for robustly learning from self-training data for low-resource sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 29–34, Minneapolis, Minnesota. Association for Computational Linguistics.

Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, Doina Caragea, and Traian Rebedea. 2022. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2711–2723, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. 2020. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *Computer Vision – ECCV 2020*, pages 171–188, Cham. Springer International Publishing.

Di Wang, Changning Tian, Xiao Liang, Lin Zhao, Lihuo He, and Quan Wang. 2024. Dual-perspective fusion network for aspect-based multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:4028–4038.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21180–21188.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5240–5252, Toronto, Canada. Association for Computational Linguistics.

Daiqing Wu, Dongbao Yang, Huawen Shen, Can Ma, and Yu Zhou. 2026. Resolving sentiment discrepancy for multimodal sentiment detection via semantics completion and decomposition. *Pattern Recognition*, 172:112719.

Daiqing Wu, Dongbao Yang, Yu Zhou, and Can Ma. 2024. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pages 5780—-5789, New York, NY, USA. Association for Computing Machinery.

Jianlong Wu, Haozhe Yang, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023. Chmatch: Contrastive hierarchical matching and robust adaptive threshold boosted semi-supervised learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15762–15772.

Wuyou Xia, Guoli Jia, Sicheng Zhao, and Jufeng Yang. 2025. Seek common ground while reserving differences: Semi-supervised image-text sentiment recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29601–29611.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2399—2402, New York, NY, USA. Association for Computing Machinery.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11525–11536. PMLR.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 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)*, pages 7617–7630, Toronto, Canada. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021a. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339, Online. Association for Computational Linguistics.

Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. 2019. Comprehensive semi-supervised multi-modal learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pages 4092—-4098. AAAI Press.

Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, and Yuan Jiang. 2021b. Semi-supervised multi-modal clustering and classification with incomplete modalities. *IEEE Transactions on Knowledge and Data Engineering*, 33(2):682–695.

Ziqi Yuan, Jingliang Fang, Hua Xu, and Kai Gao. 2024. Multimodal consistency-based teacher for semi-supervised multimodal sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3669–3683.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2023. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia*, 25:6301–6314.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: boosting semi-supervised learning with curriculum pseudo labeling. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Ming Zhang, Ke Chang, and Yunfang Wu. 2024. Multimodal semantic understanding with contrastive cross-modal feature alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11934–11943, Torino, Italia. ELRA and ICCL.

Zixing Zhang, Fabien Ringeval, Bin Dong, Eduardo Coutinho, Erik Marchi, and Björn Schüller. 2016. Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5185–5189. IEEE.

Xianbing Zhao, Xuejiao Li, Ronghuan Jiang, and Buzhou Tang. 2025. Resolving multimodal ambiguity via knowledge-injection and ambiguity learning for multimodal sentiment analysis. *Information Fusion*, 115:102745.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138, Vienna, Austria. Association for Computational Linguistics.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2023. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*, 25:3375–3385.