SABTM: Semi-Supervised Multimodal Emotion Recognition via Adaptive Barlow Twins and Mixture-of-Experts

Tiantai Zhai*
University of Electronic Science and Technology of China
Chengdu, China
202322080621@std.uestc.edu.cn

Liang Luo*
University of Electronic Science and Technology of China
Chengdu, China
luoliang@uestc.edu.cn

Yan Zhuang
University of Electronic Science and Technology of China
Chengdu, China
202211081370@std.uestc.edu.cn

Fuji Ren†
University of Electronic Science and Technology of China
Chengdu, China
renfuji@uestc.edu.cn

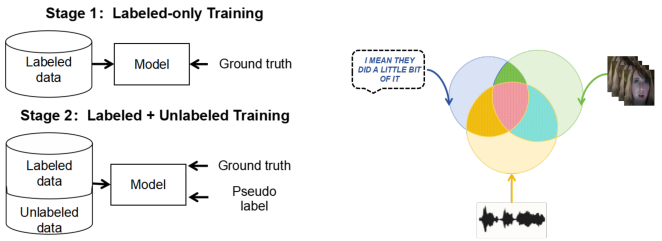
***Abstract*—Semi-supervised multimodal emotion recognition faces two critical challenges: inadequate disentanglement of shared versus modality-specific features, and modality dominance under limited labeled data. These issues lead to information loss and reduced performance. To address these problems, we propose Semi-Supervised Multimodal Emotion Recognition via Adaptive Barlow Twins and Mixture-of-Experts (SABTM). Specifically, SABTM comprises three key components: (1) Adaptive Barlow Twins (ABT) that dynamically disentangles shared and modality-specific representations, (2) Sample-wise Fusion Mixture-of-Experts (SF-MoE) module employs dynamic routing and load balancing to generate sample-specific weights, mitigating modality dominance while achieving dual-pathway fusion through feature-level and decision-level fusion to capture cross-modality interactions, and (3) Dual-Strategy Consistency Learning that enforces prediction coherence and improves pseudo-label reliability. Extensive experiments on CH-SIMS 2.0, CMU-MOSI, and CMU-MOSEI demonstrate state-of-the-art performance, with accuracy improvements of 4.62% on CH-SIMS 2.0, and 1.28% and 1.93% in 7-class accuracy on CMU-MOSI and CMU-MOSEI respectively, compared to supervised and semi-supervised learning methods.**

***Index Terms*—Semi-supervised, Multimodal, Adaptive Barlow Twins, MoE.**

I. Introduction

Multimodal Emotion Recognition (MER) integrates signals from multiple modalities to recognize human emotional states [1] [2], which has gained significant attention in both academia and industry. However, existing MER models are limited by the scale and quality of labeled datasets, and effectively leveraging unlabeled data to improve performance remains a core challenge [3].

Semi-Supervised Learning (SSL) [4] leverages both labeled and unlabeled data, typically in a two-stage manner with a labeled warm-up followed by pseudo-labeling on unlabeled samples (Fig. 1(a)). It has achieved strong results in unimodal tasks, including consistency regularization [5]. However, extending SSL to MER is not straightforward. Multimodal data inherently contains shared semantic and modality-specific information, and effective fusion should preserve both, as conceptually illustrated in Fig. 1(b). Many existing approaches rely on a single alignment strategy, using techniques like contrastive learning [6], to force representations from different modalities into a common space [7]. This alignment, while useful, tends to suppress modality-specific signals, discarding unique information that is critical for recognizing subtle emotional cues [8]. In SSL, such loss of representational quality directly reduces the reliability of pseudo-labels, further amplifying errors during training [9].



(a) Existing semi-supervised learning paradigm

(b) Information-theoretic view of shared and modality-specific features

Fig. 1. Overview of semi-supervised learning and an information-theoretic illustration of shared and modality-specific representations.

Another key limitation lies in current fusion strategies. Semi-supervised multimodal methods typically treat all modalities with a uniform fusion scheme [10] [11]. This overlooks differences in modality importance and struggles to capture nuanced interactions across modalities [12]. A promising alternative is to use multi-path predictions, which allow the model to explore richer cross-modal information. Yet this flexibility

*Tiantai Zhai and Liang Luo contributed equally to this work.
†Corresponding author.

introduces new difficulties: different prediction paths may produce inconsistent outputs. In a semi-supervised setting, such inconsistencies reduce pseudo-label stability, making it harder to distinguish reliable labels from noisy ones [13]. These conflicts ultimately degrade performance and limit the effective use of unlabeled data [4].

To address these challenges, we proposes SABTM, a novel semi-supervised framework for MER. The main contributions of this paper are summarized as follows:

- We introduce Adaptive Barlow Twins (ABT) that adaptively balances shared and modality-specific information, preserving modality-specific details while reducing information loss, leading to more expressive multimodal features.
- We develop Sample-wise Fusion Mixture-of-Experts (SF-MoE) that assigns modality importance weights at the sample level through a routing network. This enables adaptive modality fusion based on relevance to each input, addressing modality importance differences across samples while preventing modality dominance through load balancing.
- We propose a dual-strategy consistency learning module, enforcing consistency across fusion predictions to address prediction conflicts and improve pseudo-label stability and reliability in semi-supervised learning.

II. Related Work

A. Semi-Supervised Learning

Semi-supervised learning aims to leverage large amounts of unlabeled data together with limited labeled samples to improve model generalization. Early SSL methods focus on generative modeling and graph-based label propagation, while recent advances are dominated by consistency regularization and pseudo-labeling strategies [4]. Consistency-based methods enforce prediction invariance under perturbations or model stochasticity, such as Mean Teacher [9] and its variants, which have shown strong performance across vision and speech tasks. Pseudo-labeling approaches iteratively assign labels to unlabeled samples based on confident predictions, exemplified by FixMatch and its variants [14], [15].

Despite their success in unimodal settings, extending SSL to multimodal scenarios introduces additional challenges. Multimodal data exhibits heterogeneous noise levels, modality imbalance, and complex cross-modal dependencies, which can amplify pseudo-label noise when modality-specific cues are misaligned. Recent studies explore multimodal SSL via cross-modal consistency [16], [17], uncertainty-aware pseudo-label selection [13], and modality-aware interaction modeling [3].

B. Multimodal Emotion Recognition

Multimodal Emotion Recognition seeks to identify human emotional states by integrating information from multiple modalities, typically including text, audio, and visual signals [18]. To capture higher-order cross-modal interactions, a line of work has proposed explicit fusion operators. TFN [19] models unimodal, bimodal, and trimodal interactions via tensor products, while LMF [20] reduces the computational overhead of tensor fusion through low-rank factorization. Graph-based fusion has been investigated for modeling inter-modality dependencies [21]. Moreover, uncertainty-aware methods combine cross-modal attention with Bayesian inference to improve robustness under noisy multimodal signals [22].

Recent advances in MER focus on learning modality-invariant and modality-specific representations to better exploit complementary information across modalities. Methods such as MISA [23], Self-MM [7], and ConFEDE [6] explicitly decompose multimodal representations into shared and modality-specific components. Transformer-based architectures [10], [24] further enhance cross-modal modeling capacity through attention mechanisms. More recently, mixture-of-experts and uncertainty-aware routing strategies have been introduced to address modality dominance and reliability imbalance [11], [25].

III. Methodology

A. Overview

As depicted in Fig. 2, the overall architecture of SABTM follows a modular design, where ABT operates at the representation level, SF-MoE enables adaptive multimodal fusion, and the consistency module regularizes predictions across fusion pathways.

B. Problem Formulation

Semi-supervised multimodal emotion recognition aims to classify multimodal samples based on a training set consisting of labeled and unlabeled data. The labeled dataset is formally defined as $\mathcal{D}_L = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N_L}$, where each sample $\mathbf{X}_i = \{\mathbf{x}_i^v, \mathbf{x}_i^t, \mathbf{x}_i^a\}$ consists of features from visual (v), textual (t), and acoustic (a) modalities, and $y_i \in \{1, 2, \ldots, C\}$ denotes the corresponding emotion category label. The unlabeled dataset is denoted as $\mathcal{D}_U = \{\mathbf{X}_j\}_{j=1}^{N_U}$.

For each modality $m \in \{v, t, a\}$, the raw input features $\mathbf{x}^m \in \mathbb{R}^{L_m \times d_m}$ are first passed through dedicated modality encoders $\mathcal{E}^m(\cdot)$, which include pre-trained feature extraction models followed by modality-specific transformer encoders and MLP layers, to extract semantically rich hidden representations $\mathbf{h}^m = \mathcal{E}^m(\mathbf{x}^m) \in \mathbb{R}^{d_h}$, where d_h denotes the shared hidden dimension across all modalities. L_m and d_m denotes the feature length and dimension of modality m.

C. Adaptive Barlow Twins (ABT)

To address inadequate disentanglement between shared and modality-specific representations in semi-supervised multimodal emotion recognition, we propose Adaptive Barlow Twins (ABT), a decoupling mechanism designed to be robust to noisy pseudo-supervision.

Most existing disentanglement-based methods reduce cross-modal redundancy via decorrelation objectives, where the partition between shared and private subspaces is governed by a dataset-level fixed ratio [26]. Such a static design, although effective under full supervision, becomes fragile

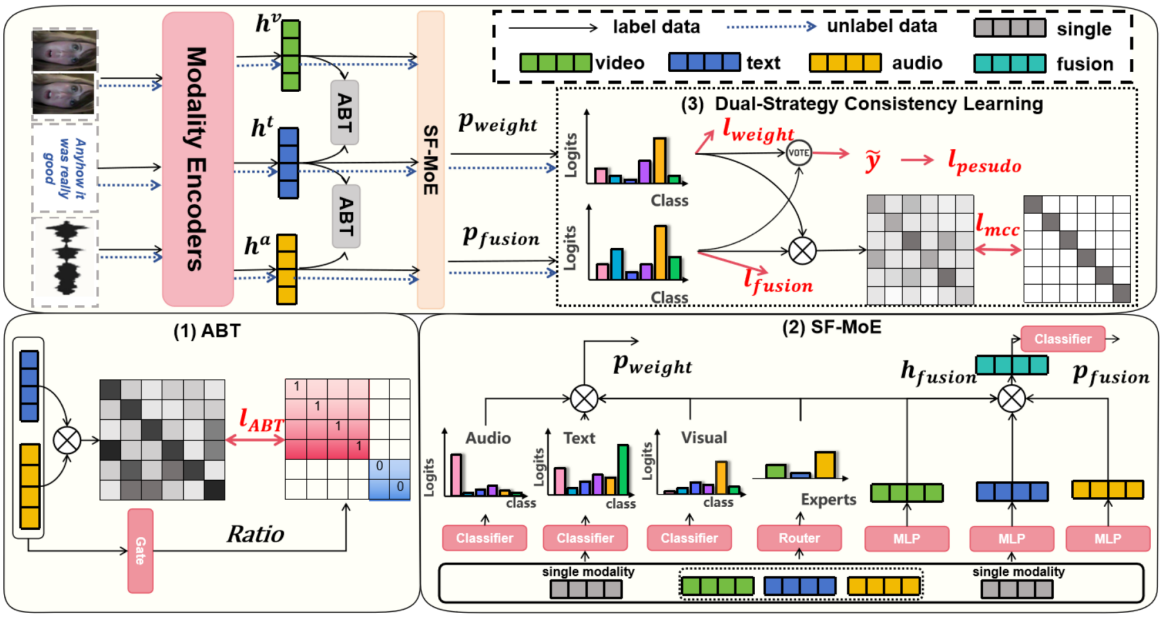


Fig. 2. Overview of the proposed SABTM with three main components that synergistically process multiple modalities: (1) Adaptive Barlow Twins (ABT) for dynamic feature decoupling, (2) Sample-wise Fusion Mixture of Experts (SF-MoE) for adaptive cross-modal fusion, and (3) Dual-Strategy Consistency Learning for coherent predictions.

in semi-supervised settings, as noisy pseudo-labels induce sample-dependent and stage-wise variations in cross-modal correlations.

ABT addresses this limitation by replacing static ratio assignment with an adaptive, correlation-aware split mechanism. It dynamically predicts the proportion of shared dimensions based on current cross-modal alignment and stabilizes the ratio evolution using an exponential moving average (EMA), eliminating the need for large-scale ratio tuning.

Specifically, for each modality pair $(p, q) \in \{(a, t), (v, t)\}$, the ABT mechanism disentangles features dynamically. Following Barlow Twins [27], batch normalization is applied to the modality representations $\mathbf{h}^m$ along the batch dimension to obtain normalized embeddings $\mathbf{z}^m$, which have zero mean and unit variance within each batch. The normalized cross-correlation matrix $\mathbf{C}^{pq} \in \mathbb{R}^{d_h \times d_h}$ between modalities p and q is computed as:

$$\mathbf{C}_{ij}^{pq} = \frac{\sum_{b=1}^{N} z_{b,i}^{p} z_{b,j}^{q}}{\sqrt{\sum_{b=1}^{N} (z_{b,i}^{p})^2} \sqrt{\sum_{b=1}^{N} (z_{b,j}^{q})^2}}. \tag{1}$$

Then a single-output MLP layer named RatioMLP is employed to predict the optimal splitting ratio for each modality pair. The ratio prediction process is formulated as:

$$r_{pq}^{raw} = \sigma(\text{RatioMLP}([\mathbf{h}^p; \mathbf{h}^q])), \tag{2}$$

where $r_{pq}^{raw} \in [0, 1]$ represents the proportion of dimensions allocated to common features, and σ denotes the sigmoid activation function.

To maintain training stability, the predicted ratios are smoothed using exponential moving averages:

$$r_{pq}^{(t)} = \theta \cdot r_{pq}^{(t-1)} + (1 - \theta) \cdot r_{pq}^{raw,(t)}, \tag{3}$$

where θ is the decay factor controlling smoothness of ratio updates, with initialization $r_{pq}^{(0)} = 0.2$ to provide a starting point for the common-unique dimension split.

Based on the stabilized ratio $r_{pq}^{(t)}$, we adaptively determine the dimensionality of shared and modality-specific subspaces as:

$$d_{h,s} = \lfloor r_{pq}^{(t)} \cdot d_h \rfloor, \quad d_{h,sp} = d_h - d_{h,s}. \tag{4}$$

Accordingly, the cross-correlation matrix $\mathbf{C}^{pq}$ can be written in a block form as

$$\mathbf{C}^{pq} = \begin{bmatrix} \mathbf{C}_{s}^{pq} & \mathbf{C}_{s \to sp}^{pq} \\ \mathbf{C}_{sp \to s}^{pq} & \mathbf{C}_{sp}^{pq} \end{bmatrix}, \tag{5}$$

where $\mathbf{C}_{s}^{pq} \in \mathbb{R}^{d_{h,s} \times d_{h,s}}$ measures cross-modal correlations within the shared subspace, $\mathbf{C}_{sp}^{pq} \in \mathbb{R}^{d_{h,sp} \times d_{h,sp}}$ corresponds to correlations within the modality-specific subspace, and $\mathbf{C}_{s \to sp}^{pq}$ and $\mathbf{C}_{sp \to s}^{pq}$ denote the off-diagonal cross-correlations between the two subspaces. In our implementation, we apply correlation constraints only to $\mathbf{C}_{s}^{pq}$ and $\mathbf{C}_{sp}^{pq}$ to stabilize training in the presence of noisy pseudo-labels.

For shared information, the shared correlation sub-matrix is encouraged to approach an identity matrix:

$$\mathcal{L}_{shr}^{pq} = \sum_{i=1}^{d_{h,s}} \left(\mathbf{C}_{s,ii}^{pq} - 1\right)^2 + \lambda \sum_{i \neq j}^{d_{h,s}} \left(\mathbf{C}_{s,ij}^{pq}\right)^2. \tag{6}$$

For modality-specific information, redundancy reduction is promoted by driving the modality-specific correlation submatrix toward zero:

$$\mathcal{L}_{sp}^{pq} = \sum_{i=1}^{d_{h,sp}} \left(\mathbf{C}_{sp,ii}^{pq}\right)^2 + \lambda \sum_{i \neq j}^{d_{h,sp}} \left(\mathbf{C}_{sp,ij}^{pq}\right)^2. \tag{7}$$

The complete ABT objective aggregates both modality pairs:

$$\mathcal{L}_{ABT} = \sum_{(p,q)\in\{(a,t),(v,t)\}} \left(\mathcal{L}_{shr}^{pq} + \mathcal{L}_{sp}^{pq}\right). \tag{8}$$

D. Sample-wise Fusion Mixture-of-Experts (SF-MoE)

To capture richer cross-modal information, we propose SF-MoE that employs a dynamic router network enabling adaptive fusion at both feature and decision levels. SF-MoE captures cross-modality interactions at feature- and decision-level and addresses sample-level modality importance differences through adaptive weighting.

Given the modality-specific representations $\mathbf{h}^v, \mathbf{h}^t, \mathbf{h}^a \in \mathbb{R}^{d_h}$ obtained from the modality encoders, the router network dynamically assesses the importance of each modality for individual samples:

$$\mathbf{w} = \text{Softmax}\left(\frac{\text{Router}([\mathbf{h}^v; \mathbf{h}^t; \mathbf{h}^a])}{\phi}\right), \tag{9}$$

where $\mathbf{w} = [w^v, w^t, w^a]$ denotes the normalized modality weights with $\sum_m w^m = 1$, and ϕ is a temperature parameter controlling the distribution sharpness. Router$(\cdot)$ is an MLP with an output dimension of 3. The softmax function transforms these outputs into a probability distribution over modalities.

At the feature level, weighted modality representations that preserve the complementary nature of multimodal information are computed through:

$$\mathbf{h}_{fusion} = \sum_{m\in\{v,t,a\}} w^m \cdot \mathbf{h}^m. \tag{10}$$

Simultaneously, at the decision level, each modality employs a dedicated classifier Classifier$_m(\cdot)$ consisting of an MLP with dropout regularization followed by a softmax layer for emotion category prediction. The individual modality predictions $\mathbf{p}^m = \text{Classifier}_m(\mathbf{h}^m)$ are then combined using the same sample-specific weights through:

$$\mathbf{p}_{weighted} = \sum_{m\in\{v,t,a\}} w^m \cdot \mathbf{p}^m. \tag{11}$$

To suppress modality dominance during fusion, we introduce a confidence-aware load balancing mechanism that aligns mixture-of-experts routing weights with modality reliability. Unlike methods relying on static modality priors, our design dynamically adjusts modality importance at the sample level and applies to both labeled and unlabeled data.

For a given sample $\mathbf{X}_i$ and modality $m \in \{v, t, a\}$, let $\mathbf{p}_i^m$ denote the predicted class probability distribution produced by the modality-specific classifier. To enable a unified formulation for semi-supervised learning, we define a unified target label $\hat{y}_i$ as

$$\hat{y}_i = \begin{cases} y_i, & \mathbf{X}_i \in \mathcal{D}_L, \\ \tilde{y}_i, & \mathbf{X}_i \in \mathcal{D}_U, \end{cases} \tag{12}$$

where $\tilde{y}_i$ denotes the pseudo-label generated by the fusion classifier:

$$\tilde{y}_i = \arg\max_{c\in\{1,\ldots,C\}} \left(\mathbf{p}_{fusion}^{(i)}\right)_c. \tag{13}$$

Based on the unified label $\hat{y}_i$, the modality-wise prediction uncertainty is defined as

$$e_i^m = 1 - p_{i,\hat{y}_i}^m, \tag{14}$$

where $p_{i,\hat{y}_i}^m$ represents the predicted probability assigned by modality m to the target class $\hat{y}_i$. This uncertainty serves as a confidence-based proxy for estimating the relative reliability of each modality without requiring ground-truth annotations for unlabeled samples.

Using the estimated uncertainties, we construct an ideal modality importance distribution $\mathbf{I}_i = [I_i^v, I_i^t, I_i^a]$ as

$$I_i^m = \frac{1/(e_i^m + \epsilon)}{\sum_{k\in\{v,t,a\}} 1/(e_i^k + \epsilon)}, \tag{15}$$

where ϵ is a small constant to ensure numerical stability. Modalities with lower uncertainty (higher confidence) are thus encouraged to receive higher importance.

To align the router-generated weights $\mathbf{w}_i = [w_i^v, w_i^t, w_i^a]$ with the ideal importance distribution while preserving expert diversity, we introduce an auxiliary load balancing loss:

$$\mathcal{L}_{aux} = \sum_m w_i^m \log(w_i^m + \epsilon) + \eta \left\|\mathbf{w}_i - \mathbf{I}_i\right\|_2^2, \tag{16}$$

where the first term promotes balanced expert utilization via entropy regularization, and the second term encourages consistency between routing decisions and modality reliability. The hyperparameter η controls the trade-off between these two objectives.

E. Dual-Strategy Consistency Learning

Although SF-MoE shows promising solution, it may generate inconsistent predictions between feature-level and decision-level fusion due to the dual-pathway design, potentially degrading pseudo-label reliability. To address this, we propose a dual-strategy consistency learning module that enforces prediction coherence across both fusion pathways.

Given the fused features $\mathbf{h}_{fusion}$ from SF-MoE, feature-level predictions $\mathbf{p}_{fusion} = \text{Classifier}_{fusion}(\mathbf{h}_{fusion})$ are derived through a MLP Classifier. To ensure consistency with decision-level predictions $\mathbf{p}_{weighted}$, a consistency constraint minimizes the distribution deviation. The consistency loss is:

$$\mathcal{L}_{mcc} = \left\|(\mathbf{p}_{fusion})^T \mathbf{p}_{weighted} - \mathbf{I}_d\right\|_F^2, \tag{17}$$

where $\mathbf{I}_d$ denotes the identity matrix.

The consistency loss promotes diagonal dominance while suppressing off-diagonal correlations, encouraging the model

to produce consistent predictions across both fusion pathways and reducing prediction uncertainty.

For reliable pseudo-label generation, the mechanism selects samples where both strategies demonstrate high confidence and prediction consistency from the unlabeled dataset $\mathcal{D}_U$. Specifically, the reliable sample set $\mathcal{R}$ includes samples where at least one pathway exceeds the confidence threshold ($\max(\mathbf{p}^{(i)}_{weighted}) > \tau$ or $\max(\mathbf{p}^{(i)}_{fusion}) > \tau$) and both pathways predict the same class label ($\arg\max(\mathbf{p}^{(i)}_{weighted}) = \arg\max(\mathbf{p}^{(i)}_{fusion})$), where τ represents the confidence threshold for pseudo-label generation. Besides, this dual-branch training employs a unified cross-entropy loss combining supervised learning on labeled samples and pseudo-supervision on reliable unlabeled samples. Thus, the training sample set becomes $\mathcal{S} = \mathcal{D}_L \cup \mathcal{R}$ with labels $\mathbf{y}^{\mathcal{S}}$ that utilize ground-truth labels y_i for labeled samples and generated pseudo-labels $\tilde{y}_i = \arg\max(\mathbf{p}^{(i)}_{fusion})$ for reliable unlabeled samples.

Since both pathways predict the same class for samples in $\mathcal{R}$, either prediction pathway suffices for pseudo-label generation. The classification loss follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\mathbf{p}^{\mathcal{S}}_{fusion}, \mathbf{y}^{\mathcal{S}}) + \mathcal{L}_{CE}(\mathbf{p}^{\mathcal{S}}_{weighted}, \mathbf{y}^{\mathcal{S}}). \quad (18)$$

F. Training Objective

The complete training objective integrates all proposed components for end-to-end optimization:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{ABT} + \mathcal{L}_{aux} + \mathcal{L}_{mcc}. \quad (19)$$

Here, the classification loss $\mathcal{L}_{cls}$ jointly optimizes shared and modality-specific representations, which effectively prevents degenerate solutions such as representation collapse when ABT is optimized together with supervised and consistency-based objectives.

IV. EXPERIMENTS

A. Datasets

To validate SABTM, we conduct comprehensive experiments on three widely-used multimodal emotion recognition benchmarks: CH-SIMS v2.0 [28], CMU-MOSI [1], and CMU-MOSEI [2]. CH-SIMS v2.0 is a Chinese in-the-wild multimodal benchmark with utterance-level sentiment/emotion labels and intensity annotations, and it additionally provides unlabeled raw video segments for semi-supervised learning. CMU-MOSI is an English dataset of opinionated video segments annotated at the utterance level for sentiment polarity and intensity. CMU-MOSEI is a larger-scale English benchmark with more diverse speakers and topics, offering a more challenging evaluation for multimodal models. Following standard benchmark protocols, Acc2 denotes binary accuracy by mapping sentiment into *negative* and *positive*, while Acc7 denotes seven-class accuracy obtained by discretizing sentiment intensity scores; F1-score is reported accordingly.

TABLE I
PERFORMANCE COMPARISON ON CH-SIMS V2.0. †: SUPERVISED LEARNING; ‡: SEMI-SUPERVISED LEARNING. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Acc2↑	F1↑
MAG-BERT [24]†	79.79	79.78
Self-MM [7]†	79.01	78.89
ConFEDE [6]†	81.88	81.84
TriagedMSA [32]†	83.75	83.67
Coupled Mamba [33]†	83.40	83.50
AV-MC [28]‡	82.50	82.55
MC-Teacher [16]‡	84.33	84.38
SABTM‡	**88.95**	**88.97**

TABLE II
PERFORMANCE COMPARISON ON CMU-MOSI AND CMU-MOSEI DATASETS. †: SUPERVISED LEARNING; ‡: SEMI-SUPERVISED LEARNING. "REPROD." DENOTES REPRODUCED RESULTS UNDER THE SAME LABELED-DATA RATIO SETTING.

Method	CMU-MOSI			CMU-MOSEI		
	Acc2↑	F1↑	Acc7↑	Acc2↑	F1↑	Acc7↑
Self-MM [7]†	85.9	85.9	45.3	85.2	85.3	53.2
ConFEDE [6]†	85.52	85.52	42.27	85.82	85.83	54.86
EUAR [11]†	86.30	86.30	46.10	86.60	86.40	54.90
TriagedMSA [32]†	86.90	87.00	–	86.30	86.40	–
SMIN [3]‡	81.55	81.45	–	86.82	86.81	–
MC-Teacher [16]‡	–	–	–	85.36	85.24	52.17
MC-Teacher (Reprod.)‡	–	–	–	86.82	86.83	55.85
SABTM‡	**87.50**	**87.56**	**47.38**	**87.62**	**87.66**	**56.79**

B. Experimental Setup

All experiments are conducted on an NVIDIA RTX 3090 GPU. We extract modality features using frozen pre-trained encoders: RoBERTa [29] for text, HuBERT [30] for audio, and CLIP [31] for visual signals. For each modality, we use the output of the last layer of the corresponding pre-trained encoder as the input features for subsequent modality encoders, and train the remaining modules end-to-end.

For semi-supervised settings, CH-SIMS v2.0 uses its labeled subset as $\mathcal{D}_L$ and the provided unlabeled raw segments as $\mathcal{D}_U$. Since CMU-MOSI and CMU-MOSEI do not provide unlabeled splits, we adopt cross-dataset unlabeled augmentation. Specifically, training on MOSI uses MOSEI utterances as $\mathcal{D}_U$ while discarding all ground-truth annotations from MOSEI; conversely, training on MOSEI uses MOSI utterances as $\mathcal{D}_U$ with MOSI labels ignored.

We optimize the model using Adam with a batch size of 256 and a learning rate of 1×10^{-5}. The hidden dimension is set to $d_h = 512$, and each modality-specific Transformer encoder contains 2 layers with 4 attention heads. We set $\eta = 0.15$ for $\mathcal{L}_{aux}$, the router temperature $\phi = 1$, and $\epsilon = 0.1$ for numerical stability. We follow a warm-up strategy by training on $\mathcal{D}_L$ for the first 15 epochs, and then enable pseudo-labeling to optimize on $\mathcal{D}_L \cup \mathcal{R}$.

C. Results

Table I shows that SABTM outperforms existing competitive state-of-the-art approaches on CH-SIMS v2.0, achieving a substantial 4.62% improvement over the previous best method

MC-Teacher. Table II presents results on CMU-MOSI and CMU-MOSEI, where SABTM achieves highly competitive performance on MOSI and superior state-of-the-art results on MOSEI across both binary and 7-class classification tasks. On CMU-MOSI, SABTM achieves a 0.6% improvement in Acc2 (87.50% vs. 86.90%) compared to TriagedMSA, while on CMU-MOSEI, SABTM shows a 0.8% improvement in Acc2 (87.62% vs. 86.82%) over SMIN. Moreover, SABTM yields consistent gains in Acc7, indicating improved fine-grained sentiment discrimination rather than merely coarse polarity separation. The larger margin on CH-SIMS further suggests that SABTM can better exploit naturally available unlabeled data, benefiting from more reliable pseudo-supervision under the proposed disentanglement, routing, and consistency mechanisms. This advantage is also reflected in improved F1 scores, suggesting that SABTM better balances precision and recall across sentiment classes. Overall, these results validate the effectiveness of SABTM's three modules and demonstrate stronger cross-modal generalization under semi-supervised training, establishing new state-of-the-art performance across all evaluated benchmarks.

D. Ablation Study

Effect of the proposed components. To validate the effectiveness of each proposed component, we conduct comprehensive ablation studies by progressively removing key modules from SABTM. Table III reports a cumulative ablation, where each row removes one additional component on top of the setting in the previous row. The performance

TABLE III
CUMULATIVE ABLATION RESULTS ON CH-SIMS V2.0. EACH ROW REMOVES ONE ADDITIONAL COMPONENT ON TOP OF THE PREVIOUS SETTING. NOTE: "W/O ROUTER" REPRESENTS USING AVERAGE WEIGHTS.

Configuration	Acc2↑	F1↑
Full Model	**88.95**	**88.97**
w/o $\mathcal{L}_{ABT}$	84.65	84.66
w/o $\mathcal{L}_{ABT}$ + w/o $\mathcal{L}_{mcc}$	82.87	82.87
w/o $\mathcal{L}_{ABT}$ + w/o $\mathcal{L}_{mcc}$ + w/o $\mathcal{L}_{aux}$	80.22	80.24
w/o $\mathcal{L}_{ABT}$ + w/o $\mathcal{L}_{mcc}$ + w/o $\mathcal{L}_{aux}$ + w/o router	82.06	82.11

degrades monotonically as more components are removed, confirming that the three modules contribute complementary benefits rather than overlapping functionality. The results demonstrate clear contributions from each component through progressive ablation. Removing ABT loss causes a 4.3% drop in Acc2, highlighting its critical role in dynamic feature disentanglement and preserving both modality-invariant and modality-specific characteristics. Further removing the mutual consistency constraint ($\mathcal{L}_{mcc}$) leads to additional degradation (82.87% Acc2), demonstrating its importance for coherent dual-pathway predictions and reliable pseudo-label generation. Progressively removing the auxiliary loss ($\mathcal{L}_{aux}$) results in the most severe degradation (80.22% Acc2), confirming its essential role in preventing modality dominance and maintaining expert diversity within the MoE framework. Finally, replacing the router network with average weights shows a 6.89% drop from the full model, validating the necessity of adaptive sample-specific modality weighting over static fusion strategies. This comparison reveals the key difference between SF-MoE with load balancing and conventional uniform weighting approaches, where dynamic weight assignment proves crucial for handling varying modality importance across samples.

TABLE IV
ABLATION ON SF-MOE FUSION PATHWAYS ON CH-SIMS V2.0. IN SINGLE-PATH VARIANTS, $\mathcal{L}_{mcc}$ IS NOT APPLICABLE BECAUSE ONLY ONE PATHWAY IS RETAINED.

Setting	$\mathcal{L}_{mcc}$	Acc2↑	F1↑
Dual-Path (Full SF-MoE)	✓	88.95	88.97
Dual-Path	–	87.99	88.00
Feature-only	–	87.31	87.31
Decision-only	–	87.04	87.01

Effect of dual-pathway design in SF-MoE. SF-MoE generates two predictions via feature-level fusion ($\mathbf{p}_{fusion}$) and decision-level fusion ($\mathbf{p}_{weighted}$). To evaluate the necessity of this design, we construct two single-path variants: *Feature-only*, which uses only feature-level fusion, and *Decision-only*, which relies on decision-level fusion. In both cases, the mutual consistency loss $\mathcal{L}_{mcc}$ is removed. Pseudo-labels are selected based on the confidence of the remaining pathway using the same threshold τ. As shown in Table IV, removing $\mathcal{L}_{mcc}$ leads to a clear drop in performance (Acc2: 88.95 $\rightarrow$ 87.99; F1: 88.97 $\rightarrow$ 88.00), indicating its role in stabilizing predictions and improving pseudo-label quality. Even without $\mathcal{L}_{mcc}$, the dual-path model still outperforms both single-path variants (87.99 vs. 87.31/87.04 Acc2), suggesting that the two fusion strategies provide complementary information. Between the single-path settings, *Feature-only* performs slightly better than *Decision-only* (87.31 vs. 87.04 Acc2), indicating that feature-level fusion is more effective for capturing cross-modal interactions in this task.

Sensitivity to ABT regularization and adaptive ratio learning. To examine the sensitivity of SABTM to key hyperparameters, we report the effect of the ABT regularization coefficient λ in Fig. 3 and compare fixed-ratio splitting with adaptive ratio learning under different EMA decay factors θ in Table V. As shown in Fig. 3, SABTM achieves the best performance when $\lambda = 0.01$. Smaller values (e.g., $\lambda < 0.01$) provide insufficient redundancy reduction, while larger values over-regularize the cross-modal correlations and degrade performance. Table V shows that adaptive ratio learning with EMA consistently outperforms fixed ratio baselines, and $\theta = 0.9$ yields the best results, indicating that moderately smoothing ratio updates improves both stability and adaptivity. Compared with the best fixed split ($r = 0.2$), ABT+EMA with $\theta = 0.9$ yields consistent gains on both MOSI and MOSEI, confirming robust adaptivity.

Qualitative analysis of feature disentanglement. To qualitatively assess whether ABT decouples shared and modality-specific representations, we visualize the learned embeddings using t-SNE. Fig. 4 presents the shared subspace (blue) and modality-specific subspaces (red/green) for two modality pairs (T–V and T–A). The modality-specific representations form

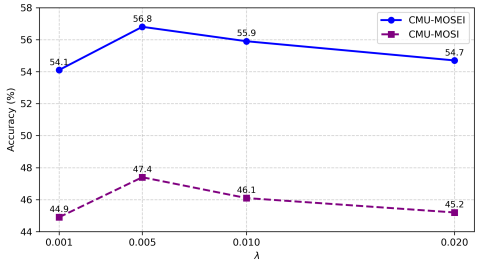


Fig. 3. Effect of the ABT regularization coefficient λ on model performance.

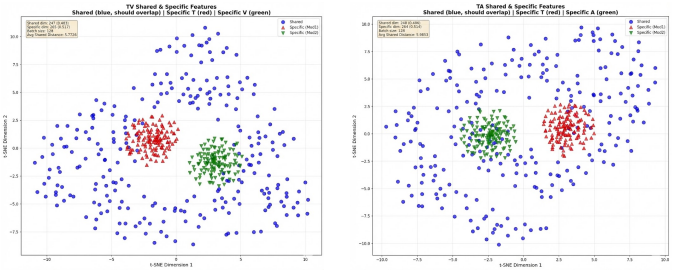


Fig. 4. t-SNE visualization of ABT-decoupled representations on a randomly sampled mini-batch. Blue denotes the modality-shared subspace, while red/green denote modality-specific subspaces. Left: T–V. Right: T–A.

TABLE V
FIXED VS. ADAPTIVE RATIO WITH EMA ON ACC7↑ (INIT=0.2 FOR ALL ADAPTIVE SETTINGS).

Setting	CMU-MOSI acc7↑	CMU-MOSEI acc7↑
Fixed $r = 0.1$	45.63	54.30
Fixed $r = 0.3$	44.90	54.86
Fixed $r = 0.2$	46.21	56.02
ABT+EMA ($\theta = 0.9$)	**47.38**	**56.79**
ABT+EMA ($\theta = 0.8$)	45.92	55.91
ABT+EMA ($\theta = 0.7$)	45.77	56.43
ABT+EMA ($\theta = 0.6$)	45.34	54.33

compact, well-separated clusters, indicating that modality-unique cues are preserved rather than absorbed into the shared space. Meanwhile, the shared representations exhibit a clear cross-modal mixing trend, suggesting that ABT captures modality-invariant semantics without forcing all information into a single cluster. Notably, a small portion of samples shows local proximity between shared and modality-specific points, which is expected under ABT's dynamic shared–specific splitting: some cues are largely shared while still retaining weak modality-dependent nuances, and may therefore lie near the boundary in the 2D projection.

Impact of pseudo-label confidence threshold. We investigate the sensitivity of SABTM to the pseudo-label confidence threshold τ, which determines the selection of reliable unlabeled samples. Table VI reports Acc7 on CMU-MOSI and CMU-MOSEI under different τ values. When τ is small, more unlabeled instances are included in the reliable set, but the increased coverage also introduces noisier pseudo-labels, which may introduce confirmation bias. Conversely, a large τ yields higher-quality pseudo-labels but selects fewer samples, reducing effective supervision from $\mathcal{D}_U$ and limiting performance. Overall, $\tau = 0.90$ achieves the best trade-off between quality and quantity and is adopted as the default setting in all experiments. Notably, a similar trend is observed on both MOSI and MOSEI, indicating that the optimal τ is not strongly dataset-dependent and generalizes well across different data scales and distributions.

Effect of Unlabeled Data (Semi-supervised vs. Fully-supervised) To quantify the contribution of unlabeled data, we conduct an additional ablation by comparing the fully-supervised setting (training only on labeled data $\mathcal{D}_L$) with the semi-supervised setting (training on $\mathcal{D}_L + \mathcal{D}_U$) under the same model architecture and optimization protocol. As shown in Table VII, introducing unlabeled samples consistently improves performance on both datasets. On CH-SIMS v2.0, semi-supervised training yields gains of $+1.91\%$ in Acc2 (88.95 vs. 87.04) and $+1.91\%$ in F1 (88.97 vs. 87.06). On CMU-MOSI, semi-supervised learning improves Acc2 by $+1.98\%$ (87.50 vs. 85.52), F1 by $+1.97\%$ (87.56 vs. 85.59), and Acc7 by $+1.17\%$ (47.38 vs. 46.21). These results demonstrate that SABTM can effectively exploit unlabeled data to enhance generalization under limited supervision, validating the necessity of the proposed semi-supervised training framework.

TABLE VI
EFFECT OF PSEUDO-LABEL THRESHOLD τ ON ACC7↑.

τ	CMU-MOSI acc7↑	CMU-MOSEI acc7↑
0.70	45.19	54.73
0.80	46.21	55.57
0.90 (default)	**47.38**	**56.79**
0.95	46.79	56.65

V. CONCLUSION

This paper investigates semi-supervised multimodal emotion recognition and proposes SABTM, an adaptive framework designed to address insufficient shared–private disentanglement and modality dominance under limited supervision. SABTM integrates three key components: Adaptive Barlow Twins to dynamically separate shared and modality-specific representations and preserve complementary cues, a Sample-wise Fusion Mixture-of-Experts module to perform sample-dependent routing with load balancing and enable dual-path fusion at both feature and decision levels, and dual-strategy consistency learning to regularize cross-path predictions for more stable training and higher-quality pseudo-labels. Extensive experiments on multiple benchmark datasets demonstrate consistent improvements over strong supervised and semi-supervised baselines, while ablation studies further confirm that each component contributes complementary benefits. In future work, we plan to develop more noise-robust semi-supervised frameworks that explicitly model and mitigate pseudo-label corruption to further improve reliability under challenging unlabeled data.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. U24A20250),

TABLE VII
COMPARISON BETWEEN FULLY-SUPERVISED AND SEMI-SUPERVISED SETTINGS ON CH-SIMS V2.0 AND CMU-MOSI (CORR REMOVED).

Training setting	CH-SIMS v2.0		CMU-MOSI		
	Acc2↑	F1↑	Acc2↑	F1↑	Acc7↑
Fully-supervised ($\mathcal{D}_L$)	87.04	87.06	85.52	85.59	46.21
Semi-supervised ($\mathcal{D}_L + \mathcal{D}_U$)	**88.95**	**88.97**	**87.50**	**87.56**	**47.38**

the Sichuan Provincial Natural Science Foundation (Grant No. 2024YFG0006 and No. 2025ZNSFSC1487), and the Fundamental Research Funds for the Central Universities (No. ZYGX2024J022 and No. ZYGX2024Z005).

REFERENCES

[1] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

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

[3] Z. Lian, B. Liu, and J. Tao, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2415–2429, 2023.

[4] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[5] Y. Xing, Q. Xu, J. Zeng, R. Huang, S. Gao, W. Xu, Y. Zhang, and W. Fan, "Cross branch feature fusion decoder for consistency regularization-based semi-supervised change detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 9341–9345.

[6] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "Confede: Contrastive feature decomposition for multimodal sentiment analysis," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 7617–7630.

[7] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[8] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.

[9] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[10] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[11] Z. Gao, D. Hu, X. Jiang, H. Lu, H. T. Shen, and X. Xu, "Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis," in *Proceedings of the 32nd ACM International Conference on Multimedia*. Melbourne VIC, Australia: Association for Computing Machinery, 2024, pp. 9650–9659.

[12] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[13] L. Wang, P. Qi, X. Bao, C. Zhou, and B. Qin, "Pseudo-label calibration semi-supervised multi-modal entity alignment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 8, 2024, pp. 9116–9124.

[14] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 596–608.

[15] C. Liu, J. Zhang, and J. Chai, "Confidence-driven semi-supervised partial label learning," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[16] Z. Yuan, J. Fang, H. Xu, and K. Gao, "Multimodal consistency-based teacher for semi-supervised multimodal sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3669–3683, 2024.

[17] J. Chen, R. Zhang, and J. Chen, "Semi-supervised multimodal classification through learning from modal and strategic complementarities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 15, 2025, pp. 15 812–15 820.

[18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[19] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2017, pp. 1103–1114.

[20] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 2247–2256.

[21] L. Maoheng, "Enhanced emotion recognition through multimodal fusion using trimodal fusion graph convolutional networks," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–9.

[22] X. Wang, M. Li, Y. Chang, X. Luo, Y. Yao, and Z. Li, "Multimodal cross-attention bayesian network for social news emotion recognition," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–9.

[23] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, 2020, pp. 1122–1131.

[24] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 2359–2369.

[25] J. Xin, S. Yun, J. Peng, I. Choi, J. L. Ballard, T. Chen, and Q. Long, "I2moe: Interpretable multimodal interaction-aware mixture-of-experts," 2025.

[26] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "Decoupling common and unique representations for multimodal self-supervised learning," in *Computer Vision – ECCV 2024: 18th European Conference*. Springer-Verlag, 2024, p. 286–303.

[27] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.

[28] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, and K. Gao, "Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module," 2022.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[30] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 3451–3460, 2021.

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

[32] Y. Luo, W. Liu, Q. Sun, S. Li, J. Li, R. Wu, and X. Tang, "Triagedmsa: Triaging sentimental disagreement in multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2025.

[33] W. Li, H. Zhou, J. Yu, Z. Song, and W. Yang, "Coupled mamba: Enhanced multimodal fusion with coupled state space model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59 808–59 832, 2024.