Multimodal Consistency-Based Teacher for Semi-Supervised Multimodal Sentiment Analysis

Ziqi Yuan, Jingliang Fang, Hua Xu, and Kai Gao

***Abstract*—Multimodal sentiment analysis holds significant importance within the realm of human-computer interaction. Due to the ease of collecting unlabeled online resources compared to the high costs associated with annotation, it becomes imperative for researchers to develop semi-supervised methods that leverage unlabeled data to enhance model performance. Existing semi-supervised approaches, particularly those applied to trivial image classification tasks, are not suitable for multimodal regression tasks due to their reliance on task-specific augmentation and thresholds designed for classification tasks. To address this limitation, we propose the Multimodal Consistency-based Teacher (MC-Teacher), which incorporates consistency-based pseudo-label technique into semi-supervised multimodal sentiment analysis. In our approach, we first propose synergistic consistency assumption which focus on the consistency among bimodal representation. Building upon this assumption, we develop a learnable filter network that autonomously learns how to identify misleading instances instead of threshold-based methods. This is achieved by leveraging both the implicit discriminant consistency on unlabeled instances and the explicit guidance on constructed training data with labeled instances. Additionally, we design the self-adaptive exponential moving average strategy to decouple the student and teacher networks, utilizing a heuristic momentum coefficient. Through both quantitative and qualitative experiments on two benchmark datasets, we demonstrate the outstanding performances of the proposed MC-Teacher approach. Furthermore, detailed analysis experiments and case studies are provided for each crucial component to intuitively elucidate the inner mechanism and further validate their effectiveness.**

***Index Terms*—Consistency-based semi-supervised learning, multimodal sentiment analysis, pseudo-label filtering.**

I. Introduction

MULTIMODAL Sentiment Analysis (MSA) is a field that aims to predict the sentiment of a speaker by analyzing various modalities such as text, audio, and visual behaviors [1], [2]. Through the advancement of sophisticated modality fusion architectures, significant progress has been made in public benchmarks [3], [4], [5]. However, these methods heavily rely on a large amount of annotated data, which can be challenging and costly to obtain [6], [7], [8]. Given the ease of collecting online video content without manual annotation compared to the time-consuming labeling process, enhancing multimodal sentiment analysis performance under the semi-supervised learning paradigm has emerged as a prominent research area. In this study, our focus is on improving MSA model performance in semi-supervised scenarios.

In recent years, a multitude of semi-supervised methodologies have emerged for traditional visual tasks, including image classification [9], [10], [11], [12], image detection [13], [14], and segmentation [15], [16]. However, the adaptation of these methods to more intricate multimodal scenarios remains a formidable challenge [17]. We summarize three fundamental challenges in the development of semi-supervised multimodal approaches. The initial challenge resides in the generalization of the consistency assumption from visual tasks to a multimodal perspective. Existing semi-supervised methods in the realm of images formulate the consistency assumption based on task-specific data augmentation, which cannot be directly extended to the multimodal domain. The second challenge revolves around the selection of a suitable multimodal backbone. Unlike conventional computer vision tasks, where default backbone networks such as ResNet [18] are commonly employed for classification, there exists no unified backbone selection for multimodal tasks. While hybrid fusion methods have achieved higher model performance, they still bears higher overfitting risks, especially when training samples are insufficient [19]. The choice of backbone architecture hinges upon the specific characteristics and requisites of the multimodal task at hand. The final challenge pertains to the effective evaluation of the value assigned to each unlabeled instance. Noisy unlabeled instances can slow down the training process or even yield detrimental

Manuscript received 6 December 2023; revised 9 April 2024; accepted 14 July 2024. Date of publication 18 July 2024; date of current version 1 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62173195, in part by the National Science and Technology Major Project towards the new generation of broadband wireless mobile communication networks of Jiangxi Province (03 and 5G Major Project of Jiangxi Province) under Grant 20232ABC03A02, in part by the High-level Scientific and Technological Innovation Talents "Double Hundred Plan" of Nanchang City in 2022 under Grant Hongke Zi (2022) 321-16, and in part by the Natural Science Foundation of Hebei Province, China under Grant F2022208006. The associate editor coordinating the review of this article and approving it for publication was Prof. Zoraida Callejas. *(Corresponding author: Hua Xu.)*

Ziqi Yuan is with the State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

Jingliang Fang is with the State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, also with the Samton (Jiangxi) Technology Development Company Ltd., Nanchang 330036, China, and also with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China.

Hua Xu is with the State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Samton (Jiangxi) Technology Development Company Ltd., Nanchang 330036, China (e-mail: xuhua@tsinghua.edu.cn).

Kai Gao is with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2024.3430543, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2024.3430543

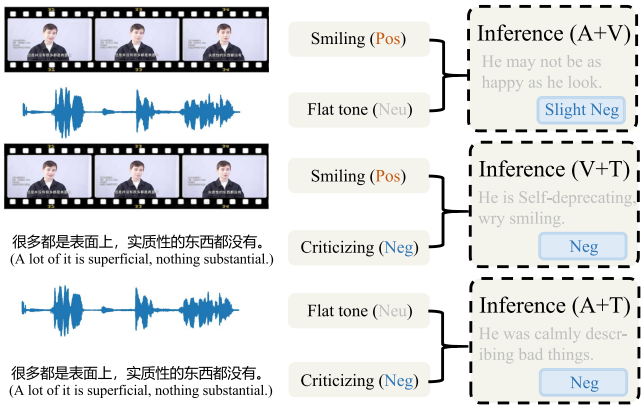

Fig. 1. Illustration of the synergistic consistency assumption, i.e. for most cases bimodal convey the same sentiment as the trimodal.

effects [20], [21]. Conventional probability threshold-based uncertainty estimation, initially proposed for image classification tasks, may prove unsuitable for the nature of multimodal sentiment analysis due to its regression characteristics. Moreover, the computational time overhead incurred by multimodal backbones imposes constraints on certain semi-supervised methods that necessitate multiple forward propagation steps for confidence calculation.

Addressing these challenges is crucial for the development of effective semi-supervised MSA methods. As for the first challenge, we propose the synergistic consistency assumption, which posits that the overall expressed sentiment of a speaker remains unchanged when any one of the text, audio, or visual modalities is dropped. Fig. 1 illustrates how the speaker's sentiment can be inferred through more than two modal behaviors, even in instances where unimodal emotions are different. For the challenge of backbone selection, we enhance the existing multitask late fusion backbone by replacing the unimodal auxiliary task with a bimodal auxiliary task. The architecture of the designed backbone is depicted in Fig. 3. This particular backbone structure is chosen for its competitive performance with a reduced number of learnable parameters, thereby ensuring computational efficiency and mitigating the risk of overfitting when training instances are limited. Building upon the aforementioned backbone, we introduce the Multimodal Consistency-based Teacher (MC-Teacher) approach, which employs a teacher-student model structure. As an improvement for sample selection compared with existing teacher-student model, we devise a novel learnable pseudo-label filtering network. Additionally, the previous exponential moving average teacher and student network are tightly coupled, resulting in the performance bottleneck since a coupled teacher is not sufficient for the student [22]. To address this concern, we design a self-adaptive coefficient for updating the teacher network, in contrast to the traditional exponential moving average method that utilizes a fixed momentum coefficient.

Experimental results show that the MC-Teacher outperforms all (semi-)supervised baselines on both text-centric and modality balanced benchmarks, exhibiting strong non-verbal behavior understanding capabilities as well as competitive text comprehension ability. To the best of our knowledge, this study represents one of the earliest works leveraging bimodal subtasks under the synergistic consistency assumption. It is also worth noting that the proposed learnable filter network exhibits excellent trade-off ability in using more unsupervised samples while maintaining the reliability of selected samples. The main contributions of this work are summarized below.

- In this work, we propose Multimodal Consistency-based Teacher (MC-Teacher), one of the earliest attempts to integrate pseudo labeling and consistency regularization into semi-supervised MSA task in order to take advantages of large amount unlabeled online video content.
- Learnable pseudo-label selection is employed, which learns how to filter misleading instances under the both implicit guidance of consistence on unsupervised data, and explicit guidance of the evaluations on supervised data instead of using fixed or adaptive threshold.
- We propose self-adaptive exponential moving average strategy, which dynamically adjusting the momentum coefficient in order to migrates the negative effects of the teacher collapsing into the students compared with trivial exponential moving average strategy.
- Extensive results demonstrate the superior performance of the proposed MC-Teacher under various Semi-Supervised Learning (SSL) settings, especially when the number of supervised data is very limited.

II. Related Works

A. Multimodal Sentiment Analysis

With the widespread adoption of sensor technology, MSA has emerged as a natural extension of previous textual sentiment analysis by incorporating acoustic and visual behaviors [23], [24], [25]. According to how to utilize information in non-verbal modalities, existing methods can be divided into two categories. The first group methods place text as centric position and integrate auxiliary audio and visual modalities into the language backbone [26], [27], [28], [29]. The literature [30] represents one of the earliest works that dynamically adjust GloVe word embeddings based on nonverbal cues. Rahman et al. [31] further extend these ideas to pretrained large language models by incorporating audio and video modal information using the multimodal adaptation gate technique. Recently, Qian et al. propose sentiment word masking as pretraining target to explicitly learn to recover text from non-verbal modality. While methods leveraging prior knowledge of text dominance can achieve competitive performance on existing datasets, they may potentially degrade in scenarios where indicative cues are present in the audio and visual modalities. The second group methods propose to treat each modality equally with the purpose of fully exploring the modality-specific behaviors in non-verbal modalities [32], [33], [34]. Literature [35], which symmetrically factorizes modalities into modality-invariant and modality-specific space, represents one of the typical efforts in this category. Yu et al. [36] also propose an symmetrical late fusion framework, which introduces auxiliary unimodal tasks for learning modality-specific cues.

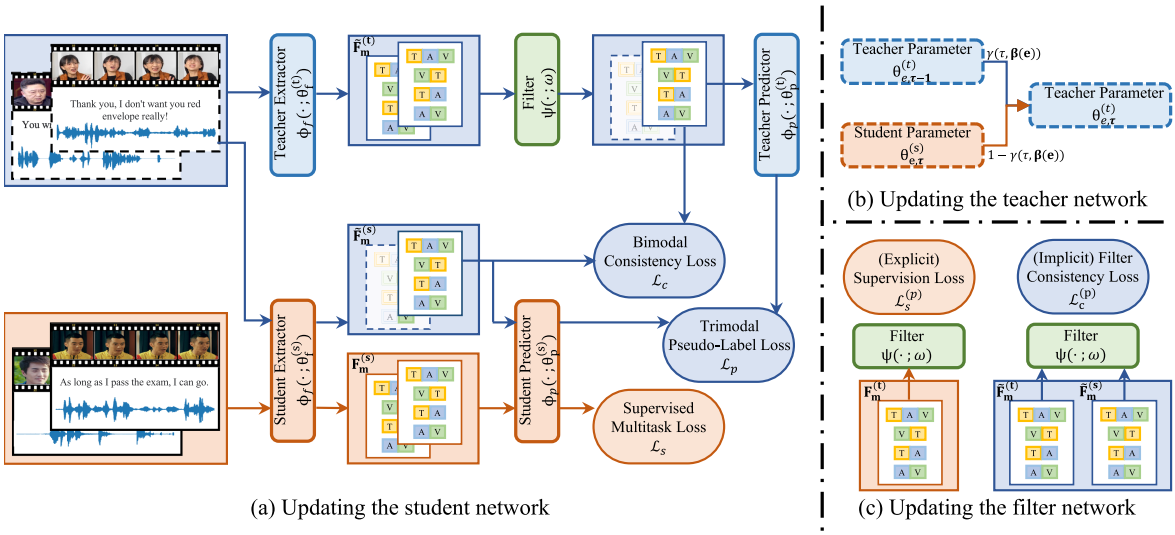

Fig. 2. Overall workflow of the Multimodal Consistency-based Teacher approach. Specifically, the workflow is segmented into three main parts, encompassing the updating of the student network (subfigure (a)), teacher network (subfigure (b)), and the filter network (subfigure (c)). The unsupervised stream is represented by the blue line, while the supervised stream is depicted by the orange line.

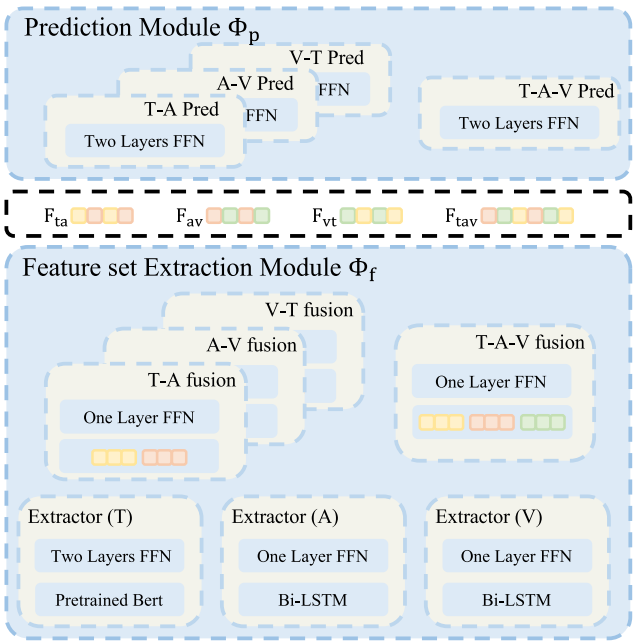

Fig. 3. Inner structure of the selected multimodal backbone, which consists of feature set extraction module Φ_f, and prediction module Φ_p.

In this study, we treat each modality equally. Furthermore, inspired from the multitask framework in [36], [37], [38], we introduce the "synergistic consistency assumption" which is an early attempt to utilize bimodal tasks instead of unimodal tasks for representation learning.

B. Consistency-Based Pseudo-Label Technique

In recent research, consistency-based pseudo-label techniques, which leverage the strengths of both consistency regularization-based and pseudo-labels-based methods, have demonstrated significant success in semi-supervised learning for computer vision tasks. Subsequent studies aim to advance these methods from two perspectives. The first category of efforts focuses on sample selection strategies to filter out misleading unlabeled data. Sohn et al. [20] first devise a fixed threshold to selectively filter low-confidence samples. Considering the different learning difficulties, class-specific thresholds technique is proposed. For generating the heuristic class-specific thresholds, Zhang et al. [11] use a curriculum learning method to take into account the learning status of each class, while Guo et al. [39] formulate the pseudo-label selection into an optimization objective by explicitly considering the number of pseudo-labels to be selected for each class. Wang et al. [21] further proposes self-adaptive threshold taking both dataset-specific and class-specific factors into account. There is also plenty of advanced efforts on adaptive threshold, such as meta learning based methods [40]. In this work, rather than using heuristic thresholds, we propose a learnable pseudo-label filtering module that directly learns how to filter error instances under both explicit supervision guidance and implicit filter consistency guidance. The second category efforts are devoted to alleviate the confirmation bias for consistency based pseudo-label methods. Ke et al. [22] decouple the teacher and student network through building two independent student network with bidirectional stabilization constraint. Xie et al. [41] propose to add noise for the student training to force student model learning harder from the pseudo labels obtained from teacher model. Liu et al. [42] designs multiple mean teachers and a student network to reduce the incorrect pseudo-label for mitigating the confirmation bias. In this work, to mitigate confirmation bias, we propose a self-adapted exponential moving average strategy to prevent the teacher network from collapsing into the student network.

C. Semi-Supervised Multimodal Sentiment Analysis

Currently, semi-supervised methods in multimodal scenarios, especially in the field of MSA, remain limited. Zhang et al. [43] are the first to employ a generative approach, utilizing a variational autoencoder to understand data distributions from both supervised and unsupervised data. Liang et al. [44] introduce an auxiliary cross-modal distribution matching task on unsupervised data to extract distinct representations for emotion recognition. Lian et al. [6] introduce a reconstruction loss based on audio and video autoencoders architecture for utilizing unsupervised data. More recently, Liu et al. [45] apply MixUp to generate unseen unimodal features and use consistency regularization on both the constructed and original features to enhance representation learning effectiveness. However, there have been few efforts to integrate the most effective consistency-based pseudo-label methods into the field of MSA. While recent studies [46], [47] suggest incorporating the vanilla pseudo-label methods, their efforts do not fully leverage the unique traits of multimodal data. In this study, we aim to bridge this gap by incorporating consistency-based pseudo-label semi-supervised learning, taking into account the characteristics of multimodal data with the proposed synergistic consistency assumption for MSA.

III. Methodology

The Multimodal Consistency-based Teacher, which builds upon the consistency-based pseudo-labeling under the synergistic consistency assumption, is a semi-supervised method tailored for multimodal scenarios. Its main novelty comes from the utilization of a learnable filter network for sample selection, as well as the incorporation of refined self-adaptive exponential moving average to decouple the teacher and student networks. This section begins by elucidating the problem statement in Section III-A, followed by providing a detailed description of the inner structure of the backbone network in Section III-B. Subsequently, the overall training workflow of the student and teacher network are introduced in Section III-C treating filter as static (Fig. 2(a) and (b)), while the description of the learnable filter network is shown in Section III-D (Fig. 2(c)). To enhance readability, Table I provides a summary of the key notations used throughout this work.

TABLE I
Table of Crucial Notations

Notations	Descriptions
$\mathrm{D}_{\mathrm{tr}}, \widetilde{\mathrm{D}}_{\mathrm{tr}}$	Supervised and unsupervised training dataset
$\mathbf{X}$, $\widetilde{\mathbf{X}}$	Supervised and unsupervised training instances
y	Ground truth sentiment annotation
m	Modality combinations, $m \in \{\mathrm{t}, \mathrm{a}, \mathrm{v}, \mathrm{ta}, \mathrm{tv}, \mathrm{va}, \mathrm{tav}\}$
$\mathbf{I}_m$	Original modality feature sequence
Φ_f	Feature set extraction module
Φ_p	Prediction module
$\theta_f^{(t)}, \theta_p^{(t)}$	Learnable parameters in teacher network
$\theta_f^{(s)}, \theta_p^{(s)}$	Learnable parameters in student network
Ψ	Filter network
ω	Learnable parameters in filter network
$\mathbf{F}_m^{(t)}(i)$	Feature of m from (t)eacher for ith labeled data
$\mathbf{F}_m^{(s)}(i)$	Feature of m from (s)tudent for ith labeled data
$\widetilde{\mathbf{F}}_m^{(t)}(i)$	Feature of m from (t)eacher for ith unlabeled data
$\widetilde{\mathbf{F}}_m^{(s)}(i)$	Feature of m from (s)tudent for ith unlabeled data
$\hat{y}_m^{(t)}(i)$	Prediction of m from (t)eacher for ith labeled data
$\hat{y}_m^{(s)}(i)$	Prediction of m from (s)tudent for ith labeled data
$\widetilde{y}_m^{(t)}(i)$	Prediction of m from (t)eacher for ith unlabeled data
$\widetilde{y}_m^{(s)}(i)$	Prediction of m from (s)tudent for ith unlabeled data
$\hat{y}_p$	Prediction of the filter network
y_p	Constructed ground truth for filter network
$\mathcal{L}_s$	Supervised loss term for student network
$\mathcal{L}_p$	pseudo-label loss for student network
$\mathcal{L}_c$	consistency regularization loss for student network
$\mathcal{L}_s^{(p)}$	Explicit prediction loss for filter network
$\mathcal{L}_c^{(p)}$	Implicit consistency loss for filter network
γ	momentum coefficient in updating the teacher network
e, τ	Epoch counts, training step counts in each epoch
α	Learning rate of the student network
β	Criteria for relative stability of the teacher network
η_m, λ_m, ξ	Crucial hyperparameters

A. Problem Statement

The task of multimodal (video) sentiment analysis is commonly formulated as a regression task. The model is trained on the designated training set denoted as $\mathrm{D}_{\mathrm{tr}} = \{(\mathbf{X}(i), y(i))\}_{i=1}^{n}$, where each instance $\mathbf{X}$ comprises three modality resources, textual token sequences, denoted as $\mathbf{I}_t \in \mathcal{R}^{L_t \times d_t}$, visual feature sequences, denoted as $\mathbf{I}_v \in \mathcal{R}^{L_v \times d_v}$, acoustic feature sequences, denoted as $\mathbf{I}_a \in \mathcal{R}^{L_a \times d_a}$, along with the scalar sentiment annotation y. In this work, we further extend the traditional MSA task into the semi-supervised scenario, where an additional set of unlabeled instances is provided, denoted as $\widetilde{\mathrm{D}}_{\mathrm{tr}} = \{(\widetilde{\mathbf{X}}(i))\}_{i=1}^{\mu n}$, with the aim of enhancing the model's performance using cost-effective online resources. The hyperparameter μ determines the relative sizes of the unlabeled instances set $\widetilde{\mathrm{D}}_{\mathrm{tr}}$ and the labeled instances set D_{tr}.

B. Inner Structure of Backbone Network

Backbone structure, i.e. the underlying structure of the teacher and student network, acts as an important role for the final model performances. As illustrated in Fig. 3, the underlying component comprises feature set extraction module and prediction module. Before we step into the inner structure of each module, we first introduce the definition of one-layer **Feed-Forward Network (FFN)** for convenience,

$$\mathrm{FFN}(\mathbf{x}) \triangleq \sigma(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}), \tag{1}$$

where σ represents the optional activation function, $\mathbf{W}$ and $\mathbf{b}$ are learnable model parameters. **Feature Set Extraction Module.** Receiving the original modality sequences $\mathbf{I}_m, m \in \{t, a, v\}$, modality specific extractor is first utilized for unimodal representation learning. For textual modality, a pre-trained BERT model [48] is used to integrate the contextual semantics. The first

time step vector, which refers to the [CLS] token[1] is then fed into two layers feed-forward network for final textual representation, which is formulated as below,

$$\mathbf{F}_t = \text{FFN}(\text{FFN}(\text{BERT}\,(\mathbf{I}_t))) \in \mathcal{R}^{d'_t}, \tag{2}$$

where $\mathbf{F}_t$ represents the extracted textual feature vector. For acoustic and visual modalities, a stack bi-directional Long Short Term Memory (LSTM) [49] followed by one layer feed-forward network is utilized as the modality encoder,

$$\mathbf{F}_a = \text{FFN}(\text{LSTM}\,(\mathbf{I}_a)) \in \mathcal{R}^{d'_a}, \tag{3}$$

$$\mathbf{F}_v = \text{FFN}(\text{LSTM}\,(\mathbf{I}_v)) \in \mathcal{R}^{d'_v}, \tag{4}$$

where $\mathbf{F}_a$ and $\mathbf{F}_v$ represents the extracted acoustic and visual feature vector separately. With the purpose of considering the consistency between bimodal and trimodal feature, we fuse each modality combination $m \subset \{t, a, v\}, |m| \geq 2$ with three individual bimodal fusion networks and one trimodal fusion network. For each modality combination $m \subset \{t, a, v\}, |m| \geq 2$, we first concatenate extracted unimodal representations followed by mapping it into a lower-dimensional space $\mathcal{R}^{d_m}$ using a one layer FFN network, taking the fusion process of visual and textual modality as an example,

$$\mathbf{F}_{\text{vt}} = \text{FFN}(\text{Concat}[\mathbf{F}_v, \mathbf{F}_t]). \tag{5}$$

The entire bimodal and trimodal feature is called a feature set. In proposed method, the teacher and student network share the same extractor architecture with individual model parameters, is formulated as $\Phi_f(\cdot; \theta_f^{(t)})$ and $\Phi_f(\cdot; \theta_f^{(s)})$ respectively.

Prediction Module: As the last step of the backbone structure, the prediction module outputs the sentiment prediction for each modality combination using a two layer FFN,

$$\hat{y}_m = \text{FFN}(\text{FFN}(\mathbf{F}_m))), \tag{6}$$

where $\hat{y}_m$ is the final sentiment prediction for modality combination $m \subset \{t, a, v\}, |m| \geq 2$. The teacher and student network also share the same predictor architecture with separate model parameters. For simplicity of further description of the workflow, we denote the teacher predictor as $\Phi_p(\cdot; \theta_p^{(t)})$ and the student predictor as $\Phi_p(\cdot; \theta_p^{(s)})$.

C. Updating the Student and Teacher Network

In this subsection, we present the simplified overview of the workflow for the student and teacher networks, with a focus on their interactions, while excluding the intricate details of the learning process of the filter network.

1) Training Process of the Student Network.: In general, as illustrated in Fig. 2(a), the devised workflow comprises two primary streams, namely the supervised stream and the unsupervised stream. In the supervised stream, when provided with labeled instances $(\mathbf{X}(i), y(i))$ from D_{tr}, we initially input the modality sequences of $\mathbf{X}(i)$ into the student network to perform bimodal and trimodal feature set extraction and sentiment prediction,

$$\mathbf{F}_m^{(s)}(i) = \Phi_f\left(\mathbf{X}(i); \theta_f^{(s)}\right),$$
$$\hat{y}_m^{(s)}(i) = \Phi_p\left(\mathbf{F}_m^{(s)}(i); \theta_p^{(s)}\right), \tag{7}$$

where $m \in \{\text{tav}, \text{ta}, \text{av}, \text{vt}\}$ refers to the bimodal and trimodal combinations. Unified sentiment annotations are utilized for providing guidance of all modality combinations, the supervised multitask loss term $\mathcal{L}_s$ is calculated as follows,

$$\mathcal{L}_s(i) = \text{L1}(\hat{y}_{\text{tav}}^{(s)}(i), y(i)) + \sum_{m \in \{\text{ta,av,vt}\}} \eta_m \cdot \text{L1}(\hat{y}_m^{(s)}(i), y(i)) \tag{8}$$

where L1 refers to the L1Loss function, and η_m represents the hyper-parameters which control the contribution of each auxiliary bimodal sentiment prediction subtask.

In the unsupervised stream, we employ trimodal pseudo-label loss and bimodal consistent regularization as guidance within the unsupervised loss term. Initially, the unlabeled data $\widetilde{\mathbf{X}}(i)$ are passed through both the student and teacher networks for feature set extraction,

$$\widetilde{\mathbf{F}}_m^{(s)}(i) = \Phi_f\left(\widetilde{\mathbf{X}}(i); \theta_f^{(s)}\right), \widetilde{\mathbf{F}}_m^{(t)}(i) = \Phi_f\left(\widetilde{\mathbf{X}}(i); \theta_f^{(t)}\right), \tag{9}$$

where $\widetilde{\mathbf{F}}_m^{(s)}(i)$ and $\widetilde{\mathbf{F}}_m^{(t)}(i)$ are the extracted feature for modality combination m from the student and teacher network separately. Receiving the extracted feature set from teacher network $\widetilde{\mathbf{F}}_m^{(t)}(i)$, the filter network is utilized with the purpose of filtering out the erroneous unlabeled instances,

$$\hat{y}_p(i) = \Psi\left(\widetilde{\mathbf{F}}^{(t)}(i); \omega\right), \tag{10}$$

where $\widetilde{\mathbf{F}}^{(t)}(i) = \text{Concat}[\widetilde{\mathbf{F}}_m^{(t)}(i)], m \in \{\text{tav}, \text{ta}, \text{av}, \text{vt}\}$ represents the concatenation of the extracted feature set from the teacher network. Additionally, $\hat{y}_p(i) \in \mathcal{R}^2$ denotes the predicted probability distribution indicating the credibility of unsupervised instance $\widetilde{\mathbf{X}}(i)$. For each selected credible unlabeled instance, the extracted feature sets from both teacher and student network are then passed to the corresponding prediction module separately,

$$\widetilde{y}_m^{(s)}(i) = \Phi_p\left(\widetilde{\mathbf{F}}_m^{(s)}(i); \theta_p^{(s)}\right), \widetilde{y}_m^{(t)}(i) = \Phi_p\left(\widetilde{\mathbf{F}}_m^{(t)}(i); \theta_p^{(t)}\right), \tag{11}$$

where $\widetilde{y}_m^{(s)}(i)$ and $\widetilde{y}_m^{(t)}(i)$ are the sentiment prediction for modality combination m from the student and teacher network respectively. For unsupervised guidance, only pseudo-label loss on trimodal sentiment prediction is utilized,

$$\mathcal{L}_p(i) = \text{L1}\left(\widetilde{y}_{\text{tav}}^{(s)}(i), \widetilde{y}_{\text{tav}}^{(t)}(i)\right). \tag{12}$$

Additionally, the bimodal consistent regularization are calculated for unsupervised guidance,

$$\mathcal{L}_c(i) = \sum_{m \in \{\text{ta,av,vt}\}} \lambda_m \cdot \text{KL}\left(\widetilde{\mathbf{F}}_m^{(s)}(i), \widetilde{\mathbf{F}}_m^{(t)}(i)\right), \tag{13}$$

where KL represents the KL divergence, which measures the similarity between the bimodal features extracted from the student and teacher networks. The hyper-parameter λ_m controls

[1]A special token in BERT language model, appended at the front of the token sequence, commonly used for sentence-level representation learning

the contribution of the bimodal consistency regularization. With the above notations, we summarize the guidance of the student network as follows,

$$\mathcal{L} = \sum_{i=1}^{B} \mathcal{L}_s(i) + \sum_{i=1}^{\mu B} \mathbf{1}\left(\arg\max \hat{y}_p(i) = 1\right)\left(\mathcal{L}_p(i) + \mathcal{L}_c(i)\right), \tag{14}$$

where B and μB represent total instance count of supervised and unsupervised instances in each training batch, respectively. $\mathbf{1}(\cdot)$ represents the indicator function, $\arg\max \hat{y}_p(i) = 1$ represents the unlabeled instance $\widetilde{\mathbf{X}}(i)$ is judged credible for providing unsupervised guidance. According to above defined loss term, the student network is updated as below,

$$\theta^{(s)} = \theta^{(s)} - \alpha \cdot \nabla_{\theta^{(s)}} \mathcal{L}, \tag{15}$$

where $\theta^{(s)} = (\theta_f^{(s)}, \theta_p^{(s)})$ refers to entire learnable parameter in student network, α is the learning rate of the student network, ∇ is the differential operator that calculates the gradient of parameters with respect to the loss function.

2) Training Process of the Teacher Network.: Obtaining the optimized the student network through gradient descent, the teacher network is then updated through the proposed self-adaptive exponential moving average strategy. In general, at τ training step in epoch e, the optimization of the teacher network follows the momentum formulation in line with the original Mean Teacher approach [50] as below,

$$\theta_{e,\tau}^{(t)} = \gamma(\tau, \beta(e))\theta_{e,\tau-1}^{(t)} + (1 - \gamma(\tau, \beta(e)))\theta_{e,\tau}^{(s)}. \tag{16}$$

Compared with the vanilla momentum updating, the proposed self-adaptive exponential moving average strategy dynamically adjust the momentum coefficient γ taking both epoch level relative stability of teacher network $\beta(e)$ as well as the training steps counts τ into consideration. Under the synergistic consistency assumption, a smaller discrepancy between the bimodal and trimodal sentiment predictions indicates a higher level of stability. The variance of the 4-dimensional prediction vector, which is the concatenation of the bimodal and trimodal predictions, is utilized to quantify such discrepancy. Therefore, we calculate the average prediction variance on all labeled and unlabeled instances (denoted as $\mathcal{V}(e)$) as the criterion for stability of the teacher network at training epoch e,

$$\mathcal{V}(e) = \frac{1}{n + \mu n}\left(\sum_{i=1}^{n} \mathrm{Var}([\hat{\mathbf{y}}^{(t)}(i)]) + \sum_{i=1}^{\mu n} \mathrm{Var}([\widetilde{\mathbf{y}}^{(t)}(i)])\right), \tag{17}$$

where Var represents variance operation, $\hat{\mathbf{y}}^{(t)}(i)$ and $\widetilde{\mathbf{y}}^{(t)}(i)$ are the concatenated prediction vectors for ith labeled instances and unlabeled instances at epoch e, respectively. The relative stability calculated by the ratio between the $\mathcal{V}(e)$ and criteria on previous epoch $\mathcal{V}(e-1)$ is utilized to control the trend of γ's changes when $e \geq 2$,

$$\beta(e) = \frac{\mathcal{V}(e-1)}{\mathcal{V}(e)}, \tag{18}$$

when $\beta(e)$ is larger the stability of the teacher network is increasing, thus we control the momentum coefficient γ to be larger to keep the teacher network unchanged. Also for each training epoch, we should gradually reduce the effect of the student network as training step τ increase. As a result the final heuristic momentum coefficient is formulated as below,

$$\gamma(\tau, \beta(e)) = \min\left(1 - \frac{1}{(1+\tau)^{\beta(e)}}, 0.97\right). \tag{19}$$

We force the momentum coefficient is less or equal than 0.97.

D. Updating the Filter Network

Filter Network: The filter network is a binary classification module which judges whether the unlabeled instance $\widetilde{\mathbf{X}}$ is credible for providing unsupervised guidance. The judgement is performed based on the extracted set of feature from the teacher network $\widetilde{\mathbf{F}}_m^{(t)}$. It is reasonable under the synergistic consistency assumption, where the filter network might learn to recognize credible instances through checking out the consistency among the bimodal and trimodal features. Specifically, the filter network composes of a simple two-layer feed forward network, which is formulated as below,

$$\Psi(\widetilde{\mathbf{F}}^{(t)}; \omega) \triangleq \mathrm{FFN}(\mathrm{FFN}(\widetilde{\mathbf{F}}^{(t)})), \tag{20}$$

where $\mathbf{F}^{(t)} = \mathrm{Concat}[\widetilde{\mathbf{F}}_m^{(t)}]$ is the concatenation of the bimodal and trimodal feature from the teacher network, and ω is the learnable parameters of the filter network, Softmax activation function is utilized in the outer feed-forward layer.

As shown in Fig. 2(b), the designed filter network is trained under both the explicit guidance of evaluation on labeled instances and the implicit guidance of consistence on unlabeled data. For labeled instances, we can determine the credible of the teacher network's prediction by explicitly comparing it with the ground truth label. An labeled instances is considered positive when both conditions are met simultaneously. Firstly the sentimental polarities of the teacher predictions and the ground truth label should be the same,

$$\mathrm{Flag}_s \triangleq (\mathrm{sgn}(\hat{y}_{tav}^{(t)}) = \mathrm{sgn}(y)) \tag{21}$$

where sgn is the signum function. Moreover, the disparity between the prediction and the annotations should not be too large. The maximum allowable difference t_d is formulated as,

$$t_d = \frac{\xi}{\tau + 1} \cdot (\mathrm{L}_{\max} - \mathrm{L}_{\min}) + \mathrm{L}_{\min}, \tag{22}$$

$$\mathrm{Flag}_d \triangleq (\mathrm{L1}(\hat{\mathrm{y}}_{tav}^{(t)}, y) < t_d) \tag{23}$$

where ξ is a hyper-parameter that controls the convergence rate of t_d, $\mathrm{L}_{\max}$ represents the difference between the strongest emotional sentiment and the neutral sentiment in the dataset, and $\mathrm{L}_{\min}$ represents the difference between each adjacent emotional category in the dataset. The constructed explicit training set is denoted as $\{(\mathbf{X}, y_p)\}_{i=1}^{B}$, where $y_p = 1$ when Flag_d and Flag_s be truth simultaneously otherwise $y_p = 0$.

Given the constructed training set, the filter network is explicitly guided by the following supervised loss,

$$\mathcal{L}_s^{(p)} = \mathrm{CELoss}\left(y_p, \Psi(\mathbf{F}^{(t)}; \omega)\right), \tag{24}$$

where CELoss refers to the cross-entropy loss, $\mathbf{F}^{(t)}$ is the concatenation of the feature set extracted from teacher network.

In addition to the explicit guidance provided by the constructed training set with supervised instances, we harness the implicit consistency loss to train the filter network. The feature set extracted from the student network can be considered as an augmentation of the teacher network. As such, the results from the filter network should maintain consistency when handling the feature sets from both the student and teacher networks for the same unlabeled instance.

$$\mathcal{L}_c^{(p)} = \mathrm{KL}\left(\Psi(\widetilde{\mathbf{F}}^{(s)};\omega), \Psi(\widetilde{\mathbf{F}}^{(t)};\omega)\right), \tag{25}$$

where KL is the KL divergence loss function, $\widetilde{\mathbf{F}}^{(t)}$ and $\widetilde{\mathbf{F}}^{(s)}$ are the concatenation of the feature set from teacher and student network of the same unsupervised instance $\widetilde{\mathbf{X}}$. The entire training pipeline is illustrated in Algorithm 1.

IV. Experimental Setups

A. Datasets

To evaluate the effectiveness of the proposed method, experiments are conducted on two benchmark MSA datasets, namely SIMS v2 [45] and CMU-MOSEI [51]. Below, we provide a concise introduction to the datasets, with comprehensive statistics detailed in the Appendix.

SIMS v2 extends the original SIMS dataset [37] with the purpose of including more instance with non-verbal affective behaviours. The dataset includes 2722/12883 labeled and unlabeled training instances, 647 validation instances, as well as 1034 testing instances. Each labeled instance contains the sentiment intensity annotation of the speaker ranging from -1 (strong negative) to +1 (strong positive).

CMU-MOSEI is one of the most common used multimodal sentiment analysis dataset that contains 16,326 utterances as training instances, 1871 utterances as validation instances, and 4659 utterances as testing instances. Each utterance is annotated with its sentiment intensity ranging from -3 (strong negative) to +3 (strong positive). In contrast to the SIMS v2, CMU-MOSEI tends to be more text-centric, with speakers often conveying their sentiments through spoken words characterized by a flat tone and inexpressive facial expressions. Given that the CMU-MOSEI dataset is fully supervised, we partition the training set into supervised set (20%) and unsupervised set (80%) to form the **CMU-MOSEI (20%)** in semi-supervised learning setting while persevering the original label distribution. Finally, the constructed **CMU-MOSEI (20%)** dataset is consist of 3266/13060 labeled and unlabeled instances for semi-supervised multimodal sentiment analysis.

B. Baseline Methods

In order to assess the effectiveness of the MC-Teacher, we perform a comparison with three tiers of baseline methods. Firstly, *supervised methods* for multimodal sentiment analysis are employed as foundational-level baselines.

TFN. Tensor Fusion Network [52] computes the outer product of the unimodal tensors to learn the fused representations.

Algorithm 1: Training Process of the MC-Teacher.

Input: labeled training set D, and unlabeled training set $\widetilde{\mathrm{D}}$.
Output: $\theta^{(s)} = (\theta_f^{(s)}, \theta_p^{(s)})$

Student network Initialization $\Phi_f(\cdot;\theta_f^{(s)})$, $\Phi_p(\cdot;\theta_p^{(s)})$.
Teacher network Initialization $\Phi_f(\cdot;\theta_f^{(t)})$, $\Phi_p(\cdot;\theta_p^{(t)})$.
Filter network Initialization $\Psi(\cdot;\omega)$.
for $e \in [1, end]$
  for $\tau \in [1, \mathrm{len}(\mathrm{dataloader}(\mathrm{D}, \widetilde{\mathrm{D}}))]$
    ▷ **Step 1: Updating the student network**
    Calculate the supervised multitask loss term $\mathcal{L}_s$ through (7) - (8).
    Calculate the trimodal pseudo-label loss term $\mathcal{L}_p$ and bimodal consistency loss term $\mathcal{L}_c$ through (9) - (13).
    Update student network $(\theta_f^{(s)}, \theta_p^{(s)})$ with (14) - (15).
    ▷ **Step 2: Updating the teacher network**
    Calculate the momentum coefficient $\gamma(\tau, \beta(e))$ with (19) and update teacher network $(\theta_f^{(t)}, \theta_p^{(t)})$ using (16).
    ▷ **Step 3: Updating the filter network (explicitly)**
    Construct supervised training set using (21) - (23), and update the filter network ω using (24).
  end for
  if $e \geq 2$
    Update the epoch-level relative stability criteria of the teacher network $\beta(e)$ using (17) - (18).
  end if
  for $\widetilde{X} \in \mathrm{dataloader}(\widetilde{\mathrm{D}})$
    ▷ **Step 4: Updating the filter network (implicitly)**
    Compute implicit consistency loss term $\mathcal{L}_c^{(p)}$ using (25), and update the filter network under the guidance of $\mathcal{L}_c^{(p)}$.
  end for
end for

LMF. The Low-rank Multimodal Fusion (LMF) [53] improves the efficiency of TFN approach through the proposed low-rank multimodal tensors fusion technique.

MulT. The Multimodal Transformer [54] extends transformer architecture fusion the source modality into the target modality using directional pairwise cross-attention mechanism.

MISA. The Modality-Invariant and -Specific Representations [35] is made up of a combination of losses including similarity loss, orthogonal loss, reconstruction loss and prediction loss to learn modality-invariant and modality-specific representation.

MAG-BERT. The Multimodal Adaptation Gate for BERT (MAG-BERT) [31] integrates auxiliary non-verbal information into the textual representations by applying multimodal adaptation gate at different layers of the BERT backbone.

Self-MM. The Self-supervised Multi-task Multimodal sentiment analysis network [36] proposes to generate the pseudo unimodal sentiment labels and then adopt them to train the model in a multi-task learning manner.

MMIM. The Multimodal InfoMax (MMIM) [28] hierarchically maximizes the mutual information in both unimodal input pairs and multimodal fusion result to maintain task related information through multimodal fusion.

CENet. The Cross-modal Enhancement Network (CENet) [26] enriches text representations by integrating visual and acoustic information into the pretrained language model.

TETFN. The Text Enhanced Transformer Fusion Network [27] learns text-oriented pairwise cross-modal mappings and generates labels for each modality to learn modality invariant and modality specific information.

ALMT. The Adaptive Language-guided Multimodal Transformer [55] uses adaptive hypermodal learning module to learn an irrelevance / conflict-suppressing representation from visual and audio features.

The *semi-supervised baselines*, characterized by their utilization of large amount unlabeled data to augment performance, serve as the secondary-tier benchmarks.

Mean Teacher. the Mean Teacher Network (Mean Teacher) [50] employs an Exponential Moving Average (EMA) teacher model to provide guidance to the student network.

VAT. The Virtual Adversarial Training Network (VAT) [56] introduces a new regularization method based on virtual adversarial loss for semi-supervised tasks.

AV-MC. The Acoustic Visual Mixup Consistent Network [45] leverages mixup technique on unlabeled instances to generate potential instances with enriched non-verbal behaviors.

MCL. The Multimodal Correlation Learning Network [57] is proposed as a supervised method, leveraging the intrinsic correlations between modalities through an auxiliary contrastive task of determining if the modality combinations come from the same video clip. In this study, we apply this contrastive learning method to semi-supervised setup.

Furthermore, the third-tier comparisons are also conducted with Claude,[2] acknowledged presently as one of the best performing *Multimodal Large Language Models (MLLM)*. In this assessment, we solely employ the provided APIs, evaluating their effectiveness within a zero-shot paradigm. Implementation details are introduced in Appendix.

C. Evaluation Metrics

In all experiments, multimodal sentiment analysis is formulated as a regression task with mean absolute error (denoted as MAE) and Correlation Coefficient (denoted as Corr) as the primary metric. In addition to these regression metrics, various classification criteria are also introduced to enable intuitive comparisons. Binary accuracy (denoted as Acc-2) and F1 score (denoted as F1) are utilized as shared classification criteria for both datasets. Furthermore, in accordance with the original dataset literature [45], [51], we reports seven-class classification accuracy (denoted as Acc-7) for the CMU-MOSEI dataset, while reports binary accuracy specifically for instances with weak positive emotions and weak negative emotions[3] (denoted as Acc-2-W) for SIMS v2, as fine-grained sentiment classification criteria. For all above metrics, higher values indicate better model performance, except for MAE, where lower values are indicative of better model performance.

[2]Specifically, claude-3-sonnet-20240229 is used for comparison purposes.

TABLE II
SELECTED CRUCIAL HYPER-PARAMETER IN THE PROPOSED MC-TEACHER APPROACH

Hyper-parameters	SIMS v2	CMU-MOSEI (20%)
$(\eta_{ta}, \eta_{av}, \eta_{vt})$	(0.4,0.2,0.8)	(0.8,0.6,0.2)
$(\lambda_{ta}, \lambda_{av}, \lambda_{vt})$	(0.4,0.6,0.6)	(0.2,0.6,1.0)
L_{max}	1	3
L_{min}	0.2	1
ξ	80	20
Student learning rate	1e-3	1e-4
Filter learning rate	0.05	0.01
Dropout	0.2	0.1

D. Experimental Details

In this study, all experiments are conducted using PyTorch library on the NVIDIA GeForce RTX 3090 with CUDA 11.4 and torch version 1.8.2. **Unaligned settings** are utilized for both datasets. For non-verbal modality feature sequence extraction, audio and visual features provided by CMU-Multimodal SDK[4] are utilized for CMU-MOSEI dataset, while audio and visual features from the SIMS v2.0 website[5] are utilized for CH-SIMS v2 dataset. Grid search based on the model performance on validation set is performed for both baselines and the proposed method. The final selected hyperparameters are recorded in Table II, while the candidate set of each hyperparameters are provided in Appendix. Additionally, for model training, an early stop strategy is employed to prevent overfitting, which halts model training if the best MAE on the validation set remains unchanged for eight consecutive epochs. To ensure fairness in comparison, all experiments are conducted three times with different random seeds, and the average performance on the testing set is reported.

V. RESULTS AND ANALYSIS

Experiments are carried out from four perspectives to verify the efficacy of the proposed MC-Teacher approach. Firstly, main experimental results containing quantitative comparison and qualitative performance curve with decreasing supervised instance counts on both CMU-MOSEI (20%) and SIMS v2 are recorded in Section V-A. Secondly, analysis of the crucial components including the loss terms (Section V-B), learnable filter network (Section V-C), and the self-adaptive EMA strategy

[3]Instances with absolute values of annotation less or equal than 0.4.
[4]https://github.com/prateekvij/CMU-MultimodalDataSDK
[5]https://thuiar.github.io/sims.github.io/chsims

TABLE III
PERFORMANCE COMPARISON ON SIMS V2 AND CMU-MOSEI (20%) DATASET

Models	SIMS v2					CMU-MOSEI (20%)				
	Corr	MAE	Acc-2	Acc-2-W	F1	Corr	MAE	Acc-2	Acc-7	F1
TFN$^{\diamond}$	65.19	0.334	77.76	69.98	77.47	71.22	0.586	77.70/82.06	50.05	78.54/82.20
LMF$^{\diamond}$	68.89	0.330	79.01	70.60	79.06	68.38	0.612	79.61/81.23	48.82	79.92/81.05
MulT$^{\diamond}$	70.32	0.317	79.50	69.61	79.59	68.93	0.599	78.86/82.20	49.77	79.39/82.12
MISA$^{\diamond}$	72.49	0.314	80.53	70.50	80.63	74.67	0.560	80.90/84.81	50.96	81.44/84.80
MAG-BERT$^{\diamond}$	69.09	0.334	79.79	71.87	79.78	73.43	0.562	80.92/84.37	50.83	81.40/84.37
Self-MM$^{\diamond}$	64.03	0.335	79.01	71.87	78.89	74.96	0.559	81.82/84.81	51.49	82.14/84.67
MMIM$^{\diamond}$	70.65	0.316	80.95	72.28	80.97	74.21	0.565	81.33/84.48	51.20	81.37/84.57
CENET$^{\diamond}$	72.21	0.298	81.53	72.67	81.58	74.63	0.559	80.77/84.31	52.09	81.37/84.57
TETFN$^{\diamond}$	70.66	0.305	80.37	71.22	80.46	74.13	0.562	80.10/83.71	51.99	80.61/83.64
ALMT$^{\diamond}$	72.05	0.312	80.65	71.84	80.77	74.93	0.577	81.22/84.26	50.03	80.84/84.38
Mean Teacher*	74.71	0.285	82.40	72.67	82.50	74.63	0.562	81.69/84.89	51.32	82.07/84.79
VAT*	74.72	0.286	82.47	73.51	82.56	74.38	0.569	81.88/84.67	50.57	82.23/84.57
AV-MC*	74.44	0.280	82.88	74.33	82.98	73.16	0.574	80.88/84.01	50.35	81.22/83.85
MCL*	73.72	0.288	83.08	75.78	83.14	73.91	0.559	81.15/84.73	52.03	81.64/84.69
Claude-3-Sonnet$^{\spadesuit}$	66.99	0.381	73.88	64.24	73.67	69.21	0.983	**84.51**/83.32	31.56	**84.06**/83.46
MC-Teacher*	**74.91**	**0.275**	**84.33**	**75.98**	**84.38**	**74.99**	**0.550**	82.98/**85.36**	**52.16**	83.25/**85.24**

"$\diamond$, $\star$" denotes fully supervised and semi-supervised MSA methods, respectively, whereas "$\spadesuit$" denotes multimodal large language model under zero-shot settings. Following the previous work [28], for ACC2 and F1 of CMU-MOSEI (20%), we have two sets of non-negative/negative (left) and positive/negative (right) evaluation results. The best results are highlighted in bold.

(Section V-D) are conducted. Thirdly, modality ablation results are in recorded in Section V-E. At last, case studies are provided in Section V-F.

A. Main Experimental Results

1) Quantitative Performance Comparison: As recorded in Table III, for a fair comparison, a quantitative evaluation is conducted on two semi-supervised datasets, namely the SIMS v2 dataset and CMU-MOSEI (20%) dataset. Given the distinct experimental settings for Claude-3-Sonnet compared to other methods, along with the presence of error instances (shown in the Appendix), we separate the analysis into two parts.

Comparison with (Semi-)Supervised Methods: According to the results, observations can be concluded from two aspects. Firstly, from model comparison aspect, it is evident that the proposed MC-Teacher surpasses all (semi-)supervised baseline methods in terms of all metrics on both datasets. Specifically, the proposed approach achieves improvements of 1.8% and 1.6% on the primary MAE metrics for the SIMS v2 and CMU-MOSEI (20%) datasets, respectively. Such result validates the effectiveness of the proposed approach. On one hand, the superior performance of MC-Teacher compared to the text-centered baseline on the text-dominated dataset CMU-MOSEI (20%) validates the competitive text comprehension ability of the proposed method. On the other hand, the exceptional performance on SIMS v2, a dataset characterized by balanced modality contribution, further demonstrates that the model can enhance its performance by effectively leveraging audio and video modality information while retaining its understanding of text modalities. Secondly, from dataset comparison aspect, the improvement of semi-supervised models on the SIMS v2 dataset is more significant than on the CMU-MOSEI (20%) dataset. Such notable performance improvement of leveraging unsupervised data on the SIMS v2 dataset can be attributed to its enriched non-verbal behaviors, which necessitate the model to encounter a wider range of modality combinations for accurate sentiment prediction.

Comparison with MLLM. As recorded in Table III, a performance comparison analysis is conducted between the proposed MC-Teacher and Claude-3-Sonnet, one of the top-performing vision-text multimodal large language models. It can be found that the proposed model outperforms Claude-3-Sonnet in all regression metrics (MAE and Corr) on both datasets. Such results indicate that the proposed method performs better on fine-grained sentiment regression tasks. Moreover, on the SIMS v2 dataset, which has clear emotional cues for audio modality, the proposed method can utilize audio resources, resulting in significantly better performance.

2) Qualitative Performance Curve: In order to further validate the semi-supervised learning capabilities of the proposed MC-Teacher, qualitative evaluations are conducted under descending counts of supervised instances. The experimental setup involves partitioning 50%, 80%, 90%, and 95% of the labeled instances in the original CMU-MOSEI (20%) and SIMS v2 datasets into unlabeled instances, while maintaining the distribution of supervised data categories. Specifically, for the CMU-MOSEI (20%) dataset, experiments are performed with supervised instances counts of $\{3266(\text{all}), 1633, 654, 327, 164\}$, while for the SIMS v2 dataset, experiments are conducted with supervised instances counts of $\{2722(\text{all}), 1361, 545, 273, 137\}$. Three baselines are selected for comparison, including one text-centric supervised method (MMIM) and two typical semi-supervised methods (AV-MC and Mean Teacher). The performance curve with descending labeled instances is depicted in Fig. 4. It can be found that the proposed MC-Teacher consistently outperforms other baselines across all

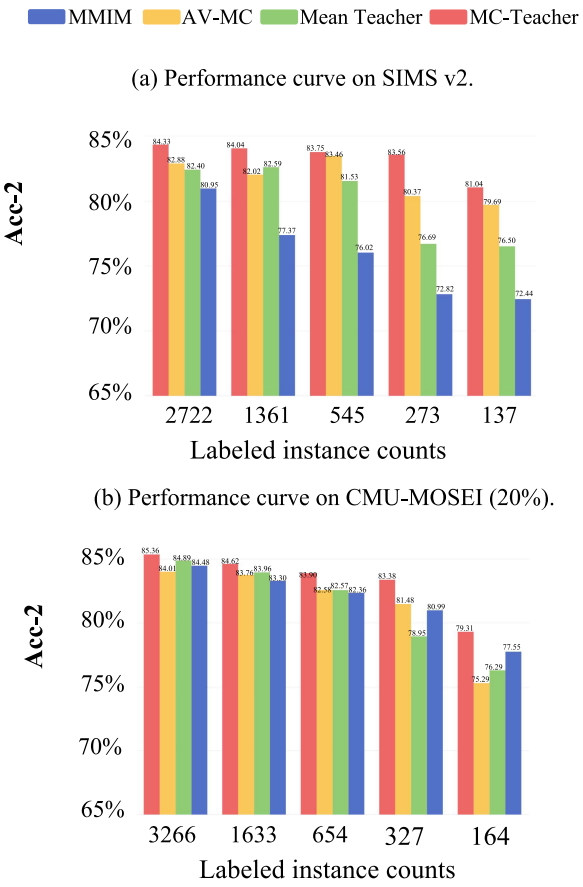

Fig. 4. Qualitative performance comparison between the proposed MC-Teacher and three typical baselines under descending labeled instance counts.

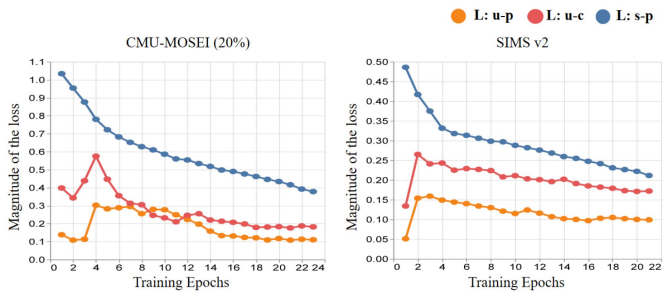

Fig. 5. Convergence analysis of the auxiliary loss term including supervised prediction loss term for bimodal tasks (L: s-p), unsupervised pseudo-label loss term (L: u-p), and unsupervised bimodal consistency loss term (L: u-c).

settings and maintains strong performance even with reduced labeled data. Specifically, when utilizing only 10% of the initial training instances on SIMS v2 and CMU-MOSEI (20%), the proposed MC-Teacher exhibits minimal performance degradation of 0.77% and 1.98%, respectively. These experimental results underscore the robustness and efficacy of the approach. Furthermore, aided by a large amount of unsupervised data, all semi-supervised methods exhibit negligible performance degradation until the training sample count decreases to 20% of the initial value on both datasets. Such result underscores the importance of further enhancing model performance by leveraging cost-effective online unlabeled instances. From dataset comparison aspect, the stability of all three semi-supervised methods on the CMU-MOSEI (20%) dataset is more significantly impacted when labeled data is decreased. This observation aligns with the earlier analysis and could be attributed to the abundance of expressive non-verbal behaviors in the SIMS v2 dataset, which facilitate model learning from unsupervised instances.

B. Analysis of the Loss Terms

In the proposed MC-Teacher approach, three auxiliary loss terms are employed alongside the primary sentiment intensity prediction loss. These include the supervised prediction loss term for bimodal tasks (denoted as L: s-p) in (8), the unsupervised pseudo-label loss term (denoted as L: u-p) in (12), and the unsupervised bimodal consistency loss term (denoted as L: u-c) in (13). In this subsection, we aim to verify the effectiveness of these three loss terms from two perspectives: ablation and convergence.

1) Ablation Studies: Ablation of the loss term is conducted on both the SIMS v2 dataset and CMU-MOSEI (20%) dataset, as shown in Table IV. Firstly, the auxiliary loss term for the supervised stream (**w/o L: s-p**) is ablated, resulting in a decrease in performance of 5.1% and 2.7% in the primary MAE criteria on the SIMS v2 and CMU-MOSEI (20%), respectively. This significant performance degradation validates the effectiveness of providing direct guidance for bimodal sentiment prediction under the synergistic consistency assumption. Secondly, the entire unsupervised stream, which includes both **L**: u-p and **L**: u-c, is ablated (denoted as **w/o L: u**). The removal of the entire unsupervised stream leads to even worse model performance compared to the **w/o L: s-p** scenario. These results provide further confirmation of the significance of utilizing unlabeled data when only a limited number of labeled instances are available. Furthermore, the pseudo-label loss term (denoted as **w/o L: u-p**) and the bimodal consistency loss term (denoted as **w/o L: u-c**) are removed separately. Similar performance is obtained when removing each component of the unsupervised stream. By comparing these scenarios with the **w/o L: u** scenario and the original MC-Teacher, it is validated that utilizing any of the unsupervised losses resulted in an enhancement of the model's performance.

2) Convergence Perspective: With the purpose of performing convergence analysis, we present the inclination of each auxiliary loss term as the training epochs progress. The obtained trajectory is depicted in Fig. 5. It is observed that the auxiliary supervised loss term L: s-p exhibits a gradual decline as the training epochs advance, eventually reaching a state of stability. Conversely, the auxiliary unsupervised loss terms (both L: u-p and L: u-c) initially experience an increase for a few training epochs, only to gradually diminish towards a stable juncture. In essence, these aforementioned patterns substantiate the convergence of both the auxiliary supervised and unsupervised loss terms. Furthermore, the increasing of the unsupervised loss terms in the first few epochs may contribute to the improvement

TABLE IV
ABLATION STUDY ON SIMS V2 AND CMU-MOSEI (20%) DATASET

Methods	SIMS v2					CMU-MOSEI (20%)				
	Corr	MAE	Acc-2	Acc-2-W	F1	Corr	MAE	Acc-2	Acc-7	F1
w/o L: s-p	74.66	0.289	81.43	71.11	81.53	74.66	0.565	80.45/84.78	51.20	81.07/84.82
w/o L: u	73.87	0.294	81.53	71.01	81.64	73.91	0.571	80.25/83.76	50.61	80.73/83.69
w/o L: u-p	**75.14**	0.278	82.98	74.33	83.06	**75.27**	0.553	82.22/84.66	51.00	82.43/84.49
w/o L: u-c	74.86	0.279	82.95	74.74	83.03	74.20	0.561	80.95/84.66	50.98	81.15/84.75
w/o filter	74.90	0.290	82.30	73.29	82.41	73.93	0.560	80.98/84.36	50.16	81.25/84.24
w/o filter & L: u-p	73.75	0.290	82.59	74.95	82.69	74.67	0.558	81.78/84.84	51.75	82.18/84.77
w/o filter & L: u-c	75.07	0.278	81.82	71.64	81.93	74.40	0.567	80.77/84.53	50.63	81.22/84.44
Ours	74.91	**0.275**	**84.33**	**75.98**	**84.38**	74.99	**0.550**	**82.98/85.36**	**52.16**	**83.25/85.24**

W/o L: s-p represents the removal of supervised loss for bimodal. **w/o L: u** represents the removal of unsupervised learning modules. **w/o L: u-p** represents the removal of the pseudo-label loss. **w/o L: u-c** represents the removal of the consistency loss. **w/o** filter represents the removal of the filter network. The best results are highlighted in bold.

of the learnable filter network, which, in its nascent stages, displays instability, thereby resulting in a substantial loss term for exceedingly incredible unlabeled instances.

C. Analysis of the Learnable Filter Network

This subsection performs analysis on the proposed learnable filter network, evaluating its impact on overall model performance as well as its inner mechanism.

1) Ablation Studies: The ablation results of the learnable filter network are recorded in the second group of the Table IV. Firstly, to evaluate the overall impact on model performance, we ablate the devised learnable filter network, which directly incorporates L: u-p and L: u-c on the entire unlabeled instances for unsupervised guidance, denoted as **w/o filter**. Under such circumstances, the model performance experiences a substantial degradation compared to the original MC-Teacher but still outperforms the **w/o L: u** scenario. This outcome underscores that leveraging unsupervised guidance on the entire unlabeled dataset can enhance model performance. Moreover, selectively filtering out misleading unlabeled instances with the proposed learnable filter network further improves the model's efficacy. Additionally, we perform ablations on the filter network along with L: u-p and L: u-c separately (employing either L: u-c or L: u-p on the entire unlabeled dataset for unsupervised guidance), denoted as **w/o filter & L: u-p** and **w/o filter & L: u-c**. It is evident that utilizing each individual loss term without the filtering process results in a degradation in performance compared to leveraging each individual loss term solely on instances identified as credible by the proposed filter network. Such above results validates the collaborative relationship between the proposed filter network and each individual unsupervised loss term.

2) Filtering Quality Analysis: Two pivotal indicators in evaluating the quality of the filtering process are considered significant: the count of selected unlabeled instances and the performance of the teacher network on the chosen unlabeled set. With the assistance of annotations from the original CMU-MOSEI dataset, we inspect the trends of these two indicators during the training process, comparing them with the **w/o filter** baseline. The corresponding curve is illustrated in Fig. 6. Across the training stages, a notable correlation between these indicators unfolds. In the initial epochs (from the 1st to the 9th), the binary accuracy of the teacher network steadily ascends, leading to a subsequent increase in the count of the selected unlabeled instances. The critical juncture transpires at the 9th epoch, wherein the filter network endeavors to judge all unlabeled instances as credible. However, this audacious attempt results in a 2% decline in the performance of the teacher network. Subsequently, the count of selected instances undergoes a reduction, stabilizing at 10,000 instances, while the binary accuracy of the teacher network on the selected unlabeled instances remains consistently around 88%. This underscores the adaptability of the filter network, facilitating swift correction of errors. Furthermore, throughout the entire training process, the performance of the teacher network on the selected unlabeled instances consistently surpasses the **w/o filter** baseline, with an average improvement exceeding 6% in each epoch. This substantiates the efficacy of the proposed filter network in effectively filtering out the misleading pseudo-labels generated by the teacher network.

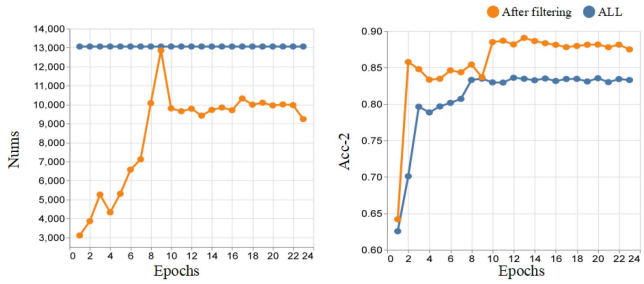

Fig. 6. The tendency of the count of the selected unlabeled instances and the binary accuracy (neg/pos classification) of the teacher network for these selected instances as the training epoch progresses.

3) Maximum Allowable Differences t_d Analysis: The maximum allowable difference t_d in (22) plays a significant role in providing explicit guidance during the training process of the filter network. We compare the proposed formulation in (22) with a fixed threshold using either $\mathrm{L}_{\min}$ or $\mathrm{L}_{\max}$, as well as a dynamic random value range from $\mathrm{L}_{\min}$ and $\mathrm{L}_{\max}$ per training

TABLE V
PERFORMANCE COMPARISON OF THE VARIOUS MAXIMUM ALLOWABLE DIFFERENCE t_d APPROACHES ON SIMS V2 AND CMU-MOSEI (20%) DATASET

Methods	SIMS v2		CMU-MOSEI (20%)	
	Acc-2	MAE	Acc-2	MAE
L_{min}	81.91	0.283	82.79/84.70	0.556
L_{max}	82.98	0.286	82.94/84.34	0.554
Rand(L_{min}, L_{max})	82.59	0.287	81.28/84.56	0.563
Ours	**84.33**	**0.275**	**82.98/85.36**	**0.550**

The best results are highlighted in bold.

TABLE VI
PERFORMANCE COMPARISON OF VARIOUS FILTERING APPROACHES

Methods	SIMS v2		CMU-MOSEI (20%)	
	Acc-2	MAE	Acc-2	MAE
Rand(25%)	82.40	0.281	80.81/84.81	0.559
Rand(50%)	81.72	0.293	78.84/84.15	0.567
Rand(75%)	82.79	0.282	79.20/84.70	0.562
MC-Drop($2e-4$)	81.62	0.290	82.01/85.06	0.554
MC-Drop($4e-4$)	80.46	0.299	80.83/84.40	0.557
MC-Drop($6e-4$)	83.37	0.287	80.02/84.23	0.558
Ours	**84.33**	**0.275**	**82.98/85.36**	**0.550**

Rand(*x*%) represents randomly preserving x% of the samples. MC-drop(*y*) represents the usage of the MC-dropout method [59] to filter samples, where only the samples with uncertainty scores below the threshold of *y* are used.

step τ. Results are presented in Table V. Notably, the model performance exhibits an average degradation of 3.8% and 1.4% on SIMS v2 and CMU-MOSEI (20%) datasets, respectively, when substituting the proposed heuristic maximum allowable differences with the other three methods. This significant degradation underscores the efficacy of the proposed maximum allowable differences in constructing explicit guidance for the training process of the filter network.

4) Filtering Strategies Analysis: With the purpose of further substantiating the efficacy of the proposed filter network, we compared it with two other filtering strategies: random filtering and the MC-Dropout approach proposed in the literature [58]. As recorded in Table VI, the quantitative results reveal that the proposed learnable filter outperforms the best substitute approach by 2.2% and 0.7% in terms of MAE for SIMS v2 and CMU-MOSEI (20%), respectively. These findings affirm the superiority of the filter network in the selection of credible instances. Additionally, it is notable that both random filtering and MC-Dropout necessitate repeated threshold adjustments to attain satisfactory performance on different datasets. In contrast, the learnable filter obviates the need for repetitive training iterations, substantially reducing the training cost of the model.

D. Analysis of the Self-Adaptive EMA Strategy

To evaluate the effectiveness of the designed self-adaptive EMA strategy, we conduct comparisons with two sets of baseline experiments for quantitative evaluation. The experimental results are outlined in Table VII. For the basic level comparison, we replace the heuristic momentum coefficient $\gamma(\tau, \beta(e))$ with random values sampled within interval approximate to the proposed self-adaptive EMA strategy ($[0.5, 1]$). This substitution led to a notable decline in model performance, with an average reduction of 2.49% in terms of binary accuracy on both datasets. For a more advanced level comparison, we replaced the proposed self-adaptive EMA strategy in MC-Teacher with the vanilla EMA strategy proposed in the literature [50]. Under this configuration, the model performance experienced a reduction of 0.57% and 1.18% in terms of binary accuracy on the SIMS v2 dataset and CMU-MOSEI (20%) dataset, respectively. All the above results underscore the superiority of the self-adaptive EMA strategy.

TABLE VII
PERFORMANCE COMPARISON OF VARIOUS TEACHER NETWORK UPDATING STRATEGY

Methods	SIMS v2		CMU-MOSEI (20%)	
	Acc-2	MAE	Acc-2	MAE
Random	81.72	0.285	81.52/85.25	0.563
EMA	83.85	0.277	82.01/85.28	0.558
Ours	**84.33**	**0.275**	**82.98/85.36**	**0.550**

Random strategy leverages random momentum coefficient per training step for updating teacher network. EMA refers to the vanilla exponential moving average strategy [50]. The best results are highlighted in bold.

TABLE VIII
MODALITY ABLATION STUDY RESULTS ON CMU-MOSEI (20%) AND SIMS V2 DATASETS

Methods	SIMS v2		CMU-MOSEI (20%)	
	Acc-2	MAE	Acc-2	MAE
(-) text	68.96	0.413	71.02/62.85	0.919
(-) audio	80.37	0.326	82.16/85.20	**0.547**
(-) vision	76.60	0.341	82.42/85.20	0.552
MC-Teacher	**84.33**	**0.275**	**82.98/85.36**	0.550

E. Analysis of the Modality Contribution

To investigate the influence of different modalities on our model's performance, we conduct a modality ablation study. This involved evaluating the model's performance under conditions where the original modality sequences are replaced with zero-padding vectors during testing. Results of these experiments are recorded in Table VIII. Analysis of the results reveals a significant impact on model performance when text is ablated across both datasets, underscoring the pivotal role of text modality in multimodal sentiment analysis. Furthermore, we noted that the impact of audio and visual modalities varies depending on the dataset. Particularly, the substantial performance decline observed when ablating audio and visual modalities on the SIMS v2 dataset highlights the importance of non-verbal emotional cues in determining speakers' sentiment, emphasizing their relevance in the analysis.

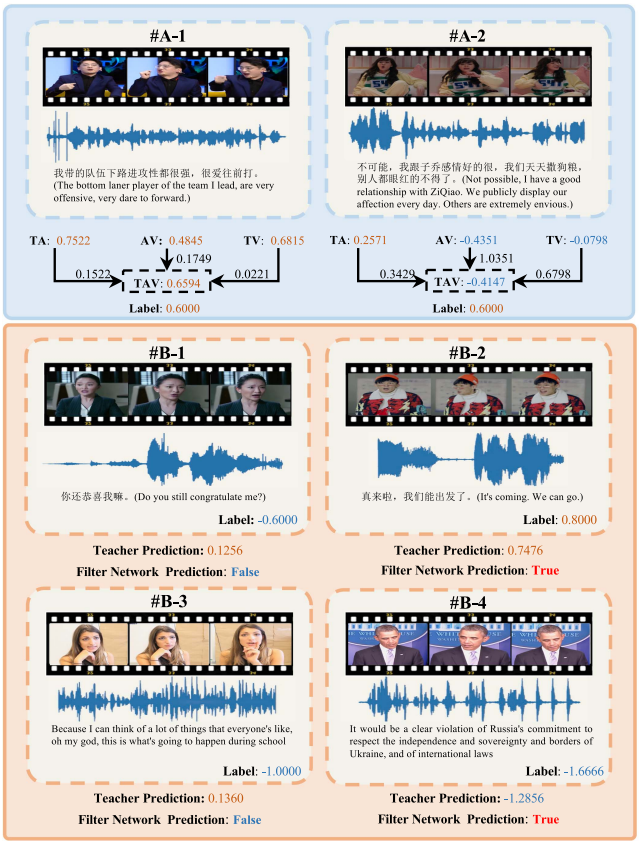

Fig. 7. Case studies. In the first group of cases within blue content, we record the bimodal, trimodal predictions and the discrepancies between them (Δ). In the second group cases within orange content, we record the predictions of the filter network, alongside the teacher prediction and the annotation.

F. Case Study

As demonstrated in Fig. 7, two groups of case studies are conducted on the testing instances within the SIMS v2.0 dataset and CMU-MOSEI (20%) dataset. The first group (instances #A-1 and #A-2) showcases the behavior of the model in scenarios where notable sentimental disparities exist in different unimodal resources. The observed better performance for the case with a more consistent bimodal prediction (instance #A-1) clearly validates the effectiveness of the proposed synergistic consistency assumption. This assumption provides a novel perspective for multimodal representation learning by leveraging the consistency among the representations of bimodal combinations, rather than relying solely on unimodal representations. The second group (instances #B-1, #B-2, #B-3, and #B-4) demonstrates the judgments of the filter network alongside the predictions of the teacher network as well as the ground truth annotations. Specifically, instance #B-1 and instance #B-2 exhibit concise textual content and vivid non-verbal cues, instance #B-3 presents a challenging example necessitating the fusion of three modalities for emotional assessment, and instance #B-4 with uninformative audio and visual behaviors centers on textual comprehension. It is evident that, in all the above scenarios, the filter network learns to filter out the unreliable instances with incorrectly predicted sentiment polarities (instance #B-1 and #B-3), based on the extracted feature set from the teacher network. These results further validate the capability of the proposed learnable filter network for the selection of unlabeled instances.

VI. Conclusion and Future Works

In this paper, our main focus is on the task of semi-supervised multimodal sentiment analysis. We leverage the unique characteristics of multimodal data and introduce the concept of synergistic consistency assumption to thoroughly explore the consistency between bimodal and trimodal representations. Building upon this assumption, we propose the Multimodal Consistency-based Teacher (MC-Teacher) approach, which represents one of the initial efforts to incorporate advanced pseudo-labeling and consistency regularization techniques into the semi-supervised multimodal sentiment analysis task. Extensive experiments are conducted to evaluate the performance of the proposed MC-Teacher approach across various semi-supervised scenarios, showcasing its remarkable efficacy. Furthermore, the analysis experiments and case studies provide intuitive additional validation of the capabilities of the learnable filter network and the self-adaptive exponential moving average strategy introduced in our method. Moreover, it is worth noticing that the proposed MC-Teacher can be easily extended to other semi-supervised multimodal tasks, making it a versatile method for enhancing model performance with cost-effective online multimedia content.

While the effectiveness of the proposed MC-Teacher approach has been verified, it still exhibits two limitations. Firstly, from multimodal understanding perspective, the proposed MC-Teacher is developed based on extracted modality feature sequences rather than directly processing the raw video as input in an end-to-end architecture. This two-phase pipeline limits certain consistency algorithms based on video augmentation techniques and inevitably leads to information loss, thereby constraining the performance of sentiment understanding. Secondly, from semi-supervision learning perspective, there is a potential drawback stemming from the inherent data waste problem of pseudo-labeled methods. Specifically, to ensure the quality of filtered samples, approximately 24% of the unsupervised data are discarded by the learnable filter (as shown in Fig. 6) and do not participate in the unsupervised learning process. In future work, addressing on above limitations, we aim to develop an end-to-end semi-supervised multimodal framework using state-of-the-art pretrained multimodal encoders and explore superior methods for generating more accurate pseudo-labels to use more (even all) unsupervised data to enhance the semi-supervised learning capability.

References

[1] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] R. Kaur and S. Kautish, "Multimodal sentiment analysis: A survey and comparison," in *Res. Anthology on Implementing Sentiment Anal. Across Mult. Disciplines*. Hershey, PA, USA: IGI Global, 2022, pp. 1846–1870.

[3] C. Chen, H. Hong, J. Guo, and B. Song, "Inter-intra modal representation augmentation with trimodal collaborative disentanglement network for multimodal sentiment analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1476–1488, 2023.

[4] J. Wu, S. Mai, and H. Hu, "Interpretable multimodal capsule fusion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1815–1826, 2022.

[5] J. Tang et al., "BAFN: Bi-direction attention based fusion network for multimodal sentiment analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1966–1978, Apr. 2023.

[6] Z. Lian, B. Liu, and J. Tao, "SMIN: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2415–2429, Jul./Sep. 2023.

[7] Z. Lian et al., "MER 2023: Multi-label learning, modality robustness, and semi-supervised learning," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 9610–9614.

[8] C. Vinola and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 2, pp. 24–44, 2015.

[9] D. Berthelot et al., "MixMatch: A holistic approach to semi-supervised learning," in *Proc. 32nd Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[11] B. Zhang et al., "FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Proc. 34th Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 18408–18419.

[12] G. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X. Hua, "Global versus localized generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1517–1525.

[13] L. Liu et al., "Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7370–7379.

[14] L. Yang et al., "Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3D object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6832–6844, Nov. 2023.

[15] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2613–2622.

[16] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.

[17] Y. Yang, K.-T. Wang, D.-C. Zhan, H. Xiong, and Y. Jiang, "Comprehensive semi-supervised multi-modal learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4092–4098.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[19] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, pp. 306–325, 2023.

[20] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. 33rd Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[21] Y. Wang et al., "FreeMatch: Self-adaptive thresholding for semi-supervised learning," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

[22] Z. Ke, D. Wang, Q. Yan, J. S. J. Ren, and R. W. H. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6727–6735.

[23] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.

[24] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[25] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[26] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Trans. Multimedia*, vol. 25, pp. 4909–4921, 2023, doi: 10.1109/TMM.2022.3183830.

[27] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognit.*, vol. 136, 2023, Art. no. 109259, doi: 10.1016/j.patcog.2022.109259.

[28] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[29] Z. Li et al., "AMOA: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 7136–7146.

[30] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.

[31] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[32] Y. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. 7th Int. Conf. Learn. Representations*, 2019.

[33] B. Li et al., "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 5923–5934.

[34] C. Jin, C. Luo, M. Yan, G. Zhao, G. Zhang, and S. Zhang, "Weakening the dominant role of text: CMOSI dataset and multimodal semantic enhancement network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 16, 2023, doi: 10.1109/TNNLS.2023.3282953.

[35] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[36] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.

[37] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[38] Q. Chen, G. Huang, and Y. Wang, "The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2689–2695, 2022.

[39] L.-Z. Guo and Y.-F. Li, "Class-imbalanced semi-supervised learning with adaptive thresholding," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 8082–8094.

[40] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11557–11568.

[41] Q. Xie, M. Luong, E. H. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10684–10695.

[42] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4248–4257. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00422

[43] D. Zhang, S. Li, Q. Zhu, and G. Zhou, "Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning," *IEEE Access*, vol. 8, pp. 22945–22954, 2020.

[44] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2852–2861.

[45] Y. Liu et al., "Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-mixup consistent module," in *Proc. Int. Conf. Multimodal Interaction*, 2022, pp. 247–258.

[46] H. Chen, C. Guo, Y. Li, P. Zhang, and D. Jiang, "Semi-supervised multi-modal emotion recognition with class-balanced pseudo-labeling," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 9556–9560.

[47] Z. Cheng et al., "Semi-supervised multimodal emotion recognition with expression mae," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 9436–9440.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.

[49] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. 28th Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[50] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 30th Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[51] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[52] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[53] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[54] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[55] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *Proc. Empirical Methods Natural Lang. Process.*, 2023, pp. 756–767.

[56] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[57] S. Mai, Y. Sun, Y. Zeng, and H. Hu, "Excavating multimodal correlation for representation learning," *Inf. Fusion*, vol. 91, pp. 542–555, 2023, doi: 10.1016/j.inffus.2022.11.003.

[58] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33nd Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059. [Online]. Available: http://proceedings.mlr.press/v48/gal16.html