Adaptive Collaborative Labeling with MLLMs for Low-Resource Multimodal Emotion Recognition

Wenwen Zhuang[1,2], Lu Xiang[1,2] * ,
Shubei Tang[1,2], Yaping Zhang[1,2], Yu Zhou[1,3]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
{zhuangwenwen2023, tangshubei2023}@ia.ac.cn, {lu.xiang, yaping.zhang, yzhou}@nlpr.ia.ac.cn

Abstract

Multimodal emotion recognition (MER) plays a crucial role in human-centric AI applications, yet existing models struggle in low-resource scenarios due to their heavy reliance on large amounts of high-quality labeled data. To address this challenge, we propose **A**daptive **C**ollaborative **L**abeling for Low-Resource MER (ACL-MER), a novel framework that leverages off-the-shelf multimodal large language models (MLLMs) to effectively exploit abundant unlabeled data. Specifically, ACL-MER incorporates a diverse teacher model zoo, wherein each MLLM specializes in a specific modality and is prompted to generate chain-of-thought predictions accompanied by scalar confidence scores. Rather than directly adopting these pseudo-labels, ACL-MER introduces an adaptive refinement strategy that selectively distills knowledge based on teacher confidence, iteratively guiding the lightweight student model toward robust learning under limited supervision. Extensive experiments on two benchmarks demonstrate that ACL-MER consistently outperforms strong baselines, especially in extremely low-resource settings.

1 Introduction

Multimodal emotion recognition (MER) is a key technology that enables machines to comprehend and respond to human emotions, with broad applications in human-computer interaction, emotional support, and intelligent education(Zhao et al., 2021; Lai et al., 2023; Yang et al., 2024). Despite its potential, the development of high-performing MER systems remains heavily dependent on large-scale, high-quality labeled datasets (Zhang et al., 2022; Shi and Huang, 2023; Zheng et al., 2023). However, the creation of such datasets is both labor-intensive and time-consuming, particularly due to the complexity and individual variability of emotional expressions (Kang and Shaver, 2004; Lian et al., 2025).

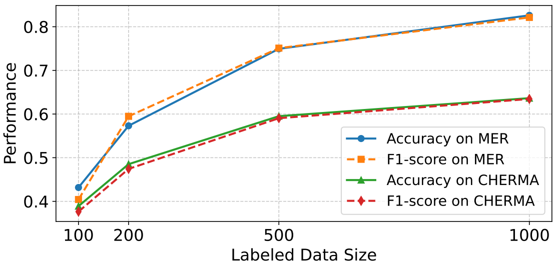


Figure 1: Performance of the model on the MER2023(Lian et al., 2023) and CHERMA(Sun et al., 2023) dataset across varying sizes of labeled training data. Both Accuracy and F1-score exhibit a substantial drop when only a few hundred labeled samples are available, indicating the limitations of pseudo-labeling in low-resource scenarios.

To reduce reliance on labeled data, pseudo-labeling(Lee et al., 2013) has become a widely adopted semi-supervised learning strategy. It starts with a small labeled seed set and trains an initial model to assign pseudo-labels to unlabeled samples, thereby effectively expanding the training set in a self-supervised manner (Cascante-Bonilla et al., 2021; Kage et al., 2024). However, when the initial labeled data is limited, the pseudo-labels tend to be noisy and unreliable, leading to error propagation and degraded performance in subsequent training iterations(Arazo et al., 2020). To illustrate this limitation, we conducted experiments using the same feature extraction methods and MER model as MERBench (Lian et al., 2024b) with varying amounts of labeled data on the MER2023 (Lian et al., 2023) and CHERMA (Sun et al., 2023) datasets. Details can be found in the Appendix B. As shown in Figure 1, performance sharply declined when the model was trained with only 100 to 200 labeled samples. This highlights a core limitation of conventional pseudo-labeling:

* Corresponding author.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2837–2853
December 20-24, 2025 ©2025 Association for Computational Linguistics

Type	Models	τ=5		τ=4	
		Proportion	Accuracy	Proportion	Accuracy
Audio LLM	Qwen2-Audio-7B (Chu et al., 2024)	11.3%	0.7872	98.5%	0.5124
Vision LLM	Qwen2-VL-7B (Wang et al., 2024a)	24.3%	0.7734	85.2%	0.3653

Table 1: Proportion–Accuracy trade-off of modality-specialized MLLMs on MER2023 (test set). For a **confidence threshold** τ, **Proportion** is the percentage of samples with confifence $c \geq \tau$, and **Accuracy** is the classification accuracy on this subset using ground-truth labels.

it struggles to produce reliable supervision when labeled data is extremely scarce.

The emergence of multimodal large language models (MLLMs) (Wu et al., 2023; Yin et al., 2023; Caffagni et al., 2024) offers new opportunities for enhancing pseudo-label generation. These models demonstrate strong perceptual and reasoning capabilities across modalities, and can generate emotion predictions in a zero-shot or few-shot manner(Hurst et al., 2024; Wang et al., 2024a; Comanici et al., 2025). In particular, recent advances have produced a variety of modality-specialized MLLMs (e.g., for audio, vision, and text), each exhibiting high competence within its domain. To explore this potential, we evaluate modality-specific MLLMs on the MER2023, prompting each model to generate emotion predictions along with scalar confidence scores. As shown in Table 1, high-confidence outputs (confidence = 5) achieve Accuracy above 0.77, outperforming small models trained on 500 labeled samples. However, these predictions only cover a limited portion of the data (e.g., 11.3% for audio, 24.3% for vision), and Accuracy degrades sharply when lower-confidence outputs are included. These observations suggest that *while MLLMs can offer valuable supervision, their coverage is limited and their reliability is confidence-dependent.*

Motivated by these findings, we introduce **ACL-MER**, a novel framework for **Adaptive Collaborative Labeling** tailored to low-resource MER scenarios. ACL-MER orchestrates a teacher model zoo composed of multiple modality-specific MLLMs. Each teacher is prompted using chain-of-thought (CoT) instructions (Wei et al., 2022), producing emotion predictions alongside scalar confidence scores. Rather than directly adopting the MLLM outputs as hard pseudo-labels, ACL-MER employs an adaptive collaboration strategy: it utilizes the confidence scores to selectively refine the probability predictions of the smaller student model. Specifically, based on the confidence scores, the outputs of the MLLMs collaboratively inform the refinement process, adjusting the student model's probability estimates. This adaptive refinement, informed by the collaborative insights of multiple MLLMs, distills knowledge from the teacher models into the smaller model, ultimately generating more robust pseudo-labels for training. Our main contributions can be summarized as follows:

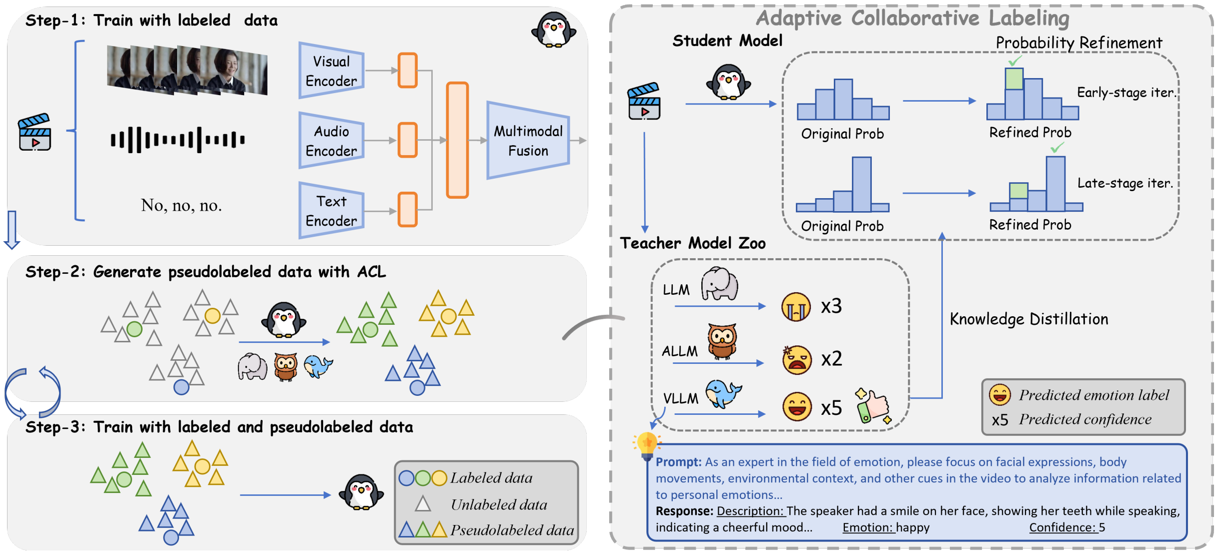
- We propose ACL-MER, an adaptive collaborative labeling framework that integrates multiple modality-specific MLLMs with a lightweight student model to enhance pseudo-label quality for low-resource MER.

- We introduce a novel adaptive collaborative mechanism that refines the student model's predictions by selectively incorporating insights from multiple MLLMs, leading to more accurate and robust pseudo-label generation.

- Sufficient experiments on MER2023 and CHERMA demonstrate that ACL-MER achieves superior performance, outperforming strong baselines in both single-teacher and conventional pseudo-labeling settings.

2 Related Work

Low-Resource Multimodal Emotion Recognition The challenge of limited labeled data in MER has driven research into various strategies. Cross-modal distillation (Albanie et al., 2018) transfers knowledge from a well-resourced modality to a less-resourced one. However, this approach necessitates labeled data in at least one modality. Liang et al. (Liang et al., 2020) leverage unlabeled data through cross-modal distribution matching, assuming consistent emotional states across modalities within utterances. Similarly, Chen et al. (Chen et al., 2023) apply a class-balanced pseudo-label approach in MER, selecting high-confidence pseudo-labels based on inter-modal classifier consistency. While these existing methods effectively exploit unlabeled data, they still depend on a certain amount of labeled data for initial training or supervision. In



Prompt: As an expert in the field of emotion, please focus on facial expressions, body movements, environmental context, and other cues in the video to analyze information related to personal emotions...

Response: Description: The speaker had a smile on her face, showing her teeth while speaking, indicating a cheerful mood... Emotion: happy Confidence: 5

Figure 2: Framework of ACL-MER. (*Left*)The iterative process of ACL-MER consists of three stages. (*Right*) The ACL module details how a Teacher Model Zoo produces confidence-scored predictions, which then refine the student model's probability distribution to generate high-quality pseudo-labels.

contrast, our proposed method demonstrates effectiveness even in extremely low-resource scenarios, with only several hundred labeled samples.

Multimodal Emotion Recognition with Large Language Models Leveraging the capabilities of MLLMs (Wu et al., 2023; Yin et al., 2023) is a promising direction for improving MER. For instance, Wang et al. (Wang et al., 2024b) prompt MLLMs to incorporate world knowledge for improved multimodal sentiment analysis. Wu et al. (Wu et al., 2024) utilize MLLMs to transform raw audio and visual data into textual emotional descriptions, thereby amplifying their emotional features. Furthermore, Lian et al. (Lian et al., 2024a) leverage MLLMs to generate diverse clues that aid in manual emotion label annotation. Other research, such as Cheng et al. (Cheng et al., 2024b) and Zhao et al. (Zhao et al., 2025), focuses on fine-tuning or reinforcement learning with MLLMs for MER tasks. In contrast, our work investigates prompting off-the-shelf MLLMs for direct emotion prediction and employs a confidence-based collaboration with a MER model to generate more reliable pseudo-labels.

3 Methodology

This section details our proposed ACL-MER. This method aims to address the performance limitations of traditional pseudo-labeling methods when labeled data is scarce.

3.1 Problem Definition

Given a small labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and a large unlabeled dataset $\mathcal{D}_u = \{x_i\}_{i=1}^{N_u}$, our goal is to train a robust MER model. Each sample x_i comprises multimodal information from acoustic (x_i^a), visual (x_i^v), and textual (x_i^t) modalities. The emotion label y_i is drawn from a predefined label space $y_i \in \{1, 2, \ldots, C\}$. We assume $|\mathcal{D}_u| \gg |\mathcal{D}_l|$, characterizing a low-resource setting.

3.2 Overview of ACL-MER

As illustrated in Figure 2, ACL-MER iteratively enhances a student model through an adaptive collaboration framework with MLLMs. Each iteration includes: (1) training a small model on the current labeled data, (2) refining the model's probability distribution using MLLM confidence scores to generate pseudo-labels, and (3) retraining with high-confidence pseudo-labels. This iterative process, with model re-initialization, continues until a stopping criterion is met, mitigating concept drift (Cascante-Bonilla et al., 2021).

3.3 Initial Student Model Training

We begin the training process by training a student model on the limited labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$. The student model encompasses three key components: feature extraction, multimodal fusion, and model optimization.

Feature Extraction. For each input x_i, we extract visual (F_i^v), acoustic (F_i^a), and textual (F_i^t)

features using modality-specific pre-trained encoders:

$$F_i^m = \text{Encoder}_m(x_i^m) \in \mathbb{R}^{d_m},\ m \in \{v, a, t\} \tag{1}$$

where d_v, d_a, and d_t are the corresponding feature dimensions.

Multimodal Fusion. The extracted features are then fused using a multimodal fusion module:

$$F_i = \text{MultimodalFusion}(F_i^v, F_i^a, F_i^t) \in \mathbb{R}^{d_f} \tag{2}$$

$$\hat{y}_i = \text{Softmax}(\text{MLP}(F_i)) \in \mathbb{R}^C \tag{3}$$

We explore various fusion models in our experiments.

Training Objective. The student model is trained by minimizing the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N_l}\sum_{i=1}^{N_l}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{4}$$

where $\hat{y}_{ic}$ is the predicted probability and y_{ic} is the ground-truth label. This provides a basis for generating pseudo-labels in the subsequent stages.

3.4 Adaptive Collaborative Labeling

The module of Adaptive Collaborative Labeling (ACL) is the core component of our approach, designed to generate high-quality pseudo-labels by leveraging the diverse knowledge and reasoning capabilities of multiple MLLMs. This module operates iteratively, refining the student model's predictions through collaborative knowledge distillation.

Teacher Model Zoo. The ACL module employs a teacher model zoo $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$, where each T_m represents a different MLLM. These models can vary in architecture, training data, and modalities they emphasize.

Prompt-Based Inference with MLLMs. For each teacher model T_m, a suitable prompt P_i is manually designed. By employing chain-of-thought prompting, we aim to elicit step-by-step reasoning from the model, enabling it to produce output labels accompanied by their associated confidence scores. The output of each teacher model T_m can be represented as:

$$O_{im} = T_m(P_i(\tilde{x}_i)) = (y_{im}, c_{im}) \tag{5}$$

where y_{im} is the predicted emotion label and c_{im} is a self-estimated confidence score ranging from 1 to 5. Since the output of each teacher model depends solely on the input sample, such inference is performed only once, thereby ensuring the efficiency of our method.

Probability Refinement. Given student model probabilities $P_s(\hat{y}_i|\tilde{x}_i)$, we refine it based on the teacher's confidence. We introduce a confidence threshold c_t and a probability adjustment value p_t. If $c_{im} > c_t$, we adjust the probability of the predicted label y_{im}. The refinement process can be formulated as:

$$P_s'(\hat{y}_i|\tilde{x}_i) = \begin{cases} P_s(\hat{y}_i|\tilde{x}_i) + p_t, & \text{if } y = y_{im} \text{ and } c_{im} > c_t \\ P_s(\hat{y}_i|\tilde{x}_i), & \text{otherwise} \end{cases} \tag{6}$$

where $P_s'(\hat{y}_i|\tilde{x}_i)$ is the refined probability distribution.

The p_t is dynamically adjusted across iterations (t). This adjustment aims to gradually reduce the influence of the teacher models as the student learns and becomes more proficient.Specifically, p_t is defined as:

$$p_t = \max(p_{init} - \gamma \cdot t, p_{min}) \tag{7}$$

where p_{init}, γ, and p_{min} are initial adjustment, decay rate, and lower bound, respectively.

Pseudo-Label Selection. After refinement, we select high-confidence predictions as pseudo-labels. A sample $\tilde{x}_i$ is selected if:

$$\max_y P_s'(\hat{y}_i|\tilde{x}_i) > \theta \tag{8}$$

where θ is a selection threshold. To mitigate class imbalance, we select up to *top_k* samples per class y with highest $P_s'(\hat{y}_i|\tilde{x}_i)$ values. The selected samples form the set $\mathcal{L}$.

Dataset Update. The labeled dataset $\mathcal{D}_l$ and unlabeled dataset $\mathcal{D}_u$ are updated using the selected pseudo-labeled data $\mathcal{L}$. Specifically, $\mathcal{L}$ is moved from $\mathcal{D}_u$ to $\mathcal{D}_l$, represented as:

$$\mathcal{D}_l^{t+1} = \mathcal{D}_l^t \cup \mathcal{L} \tag{9}$$

$$\mathcal{D}_u^{t+1} = \mathcal{D}_u^t \setminus \mathcal{L} \tag{10}$$

This iterative process allows the student model to learn from increasingly larger and more reliable sets of labeled data, thereby improving its performance over time.

MER2023				CHERMA			
Emotion	**Train&Val**	**Test**	**Unlabeled**	**Emotion**	**Train&Val**	**Test**	**Unlabeled**
neutral	259	166	-	neutral	211	1216	-
angry	244	183	-	angry	208	1128	-
happy	204	169	-	sad	181	1002	-
worried	137	45	-	happy	123	715	-
sad	113	257	-	disgust	112	633	-
surprise	43	14	-	surprise	98	552	-
-	-	-	-	fear	67	420	-
Total	**1000**	**834**	**73148**		**1000**	**5666**	**15918**

Table 2: Statistics of data samples for MER2023 and CHERMA dataset.

4 Experiments

4.1 Experimental Setup

Datasets We conducted experiments using two widely-used Chinese multimodal emotion recognition datasets: MER2023 (Lian et al., 2023) and CHERMA(Sun et al., 2023). Dataset statistics are detailed in Table 2. Low-resource scenarios were simulated by sampling 100, 200, 500, and 1000 instances from the original Train sets, maintaining class distributions. For CHERMA, unused training data was treated as unlabeled.

Evaluation Metrics We use the weighted average F1-score and Accuracy as evaluation metrics(Powers, 2020). The weighted average F1-score better reflects the model's performance on imbalanced datasets, while Accuracy measures the model's overall classification accuracy.

Implementation Details Audio, text, and visual features were extracted using HUBERT-large (Hsu et al., 2021), RoBERTa-large (Liu et al., 2019), and CLIP-large (Radford et al., 2021) models, respectively. Qwen2-Audio-7B-Instruct (Chu et al., 2024), Qwen2-VL-7B-Instruct (Wang et al., 2024a), and Qwen2.5-7B-Instruct (Team, 2024) were used as teacher models. The hyperparameters were set as follows: $p_{min} = 0.2$, $p_{init} = 0.6$, $\gamma = 0.1$, $\theta = 0.995$, and $top\_k = 200$. All experiments were conducted on a Tesla V100 GPU. Further details can be found in the Appendix D.

Baselines To thoroughly evaluate ACL-MER, we conduct experiments using several established base models, including a standard attention mechanism (Vaswani, 2017), MMIM (Han et al., 2021), and LMF (Liu et al., 2018). More detailed descriptions of these works can be found in the Appendix C.

Furthermore, we compare our method against various pseudo-labeling strategies,with all hyperparameters matching those of ACL-MER:

- **No Pseudo-Labeling(No PL):** This baseline trains the model exclusively on the available labeled data, without incorporating any unlabeled samples.
- **Teacher-only Pseudo-Labeling(T-PL):** This approach directly uses predictions from MLLMs as pseudo-labels for unlabeled data.
- **Student-only Pseudo-Labeling(S-PL):** This baseline employs a student model to generate pseudo-labels for unlabeled data, specifically using the class-balanced pseudo-labeling method (Chen et al., 2023).

4.2 Main Results

Table 3 presents the performance of our proposed ACL-MER compared to various baselines under different levels of labeled data on both the MER2023 and CHERMA datasets.

Overall Performance of ACL-MER Our proposed ACL-MER consistently achieves superior performance across all experimental settings, outperforming the baselines in both F1-score and Accuracy. This highlights the effectiveness of our adaptive collaborative labeling framework, particularly in low-resource scenarios. For instance, on the MER2023 dataset with only 100 labeled samples(line 4), ACL-MER with the Attention base model achieves an F1-score of 0.6753 and an Accuracy of 0.6775, significantly surpassing other methods. Similar trends are observed for other base models and on CHERMA dataset, demonstrating the robustness and generalizability of ACL-MER.

Comparison with Pseudo-Labeling Strategies Limitations of No PL: As anticipated, training exclusively on limited labeled data (No PL) resulted in the lowest performance, highlighting the critical need for semi-supervised learning in low-resource MER.

No.	Base Model	Method	$\mathrm{N}_{\mathrm{labeled}}$=100		$\mathrm{N}_{\mathrm{labeled}}$=200		$\mathrm{N}_{\mathrm{labeled}}$=500		$\mathrm{N}_{\mathrm{labeled}}$=1000	
			F1-score	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**
					Dataset: MER2023					
1	Attention	No PL	0.4042	0.4317	0.5948	0.5731	0.7510	0.7494	0.8213	0.8261
2		T-PL	0.6626	0.6738	0.7107	0.7206	0.7659	0.7734	0.8094	0.8189
3		S-PL	0.5320	0.5432	0.6868	0.6679	0.7692	0.7698	0.8368	0.8393
4		ACL-MER	**0.6753**	**0.6775**	**0.7672**	**0.7674**	**0.8255**	**0.8273**	**0.8416**	**0.8489**
5	MMIM	No PL	0.3994	0.4185	0.5114	0.4964	0.7529	0.7494	0.8085	0.8129
6		T-PL	0.5578	0.5679	0.6690	0.6734	0.7558	0.7674	0.7921	0.8022
7		S-PL	0.3389	0.4029	0.6160	0.5851	0.7739	0.7734	0.8032	0.8070
8		ACL-MER	**0.5713**	**0.6007**	**0.6805**	**0.6847**	**0.8183**	**0.8237**	**0.8339**	**0.8417**
9	LMF	No PL	0.3344	0.3729	0.5425	0.5312	0.7825	0.7830	0.8215	0.8213
10		T-PL	0.4914	0.5022	0.5122	0.5266	0.7691	0.7758	0.8040	0.8118
11		S-PL	0.4468	0.4796	0.5541	0.5528	0.7823	0.7926	0.8171	0.8249
12		ACL-MER	**0.4867**	**0.5120**	**0.5556**	**0.5612**	**0.8018**	**0.8141**	**0.8255**	**0.8321**
					Dataset: CHERMA					
13	Attention	No PL	0.3764	0.3890	0.4743	0.4852	0.5905	0.5951	0.6345	0.6364
14		T-PL	0.5243	0.5451	0.5311	0.5605	0.5686	0.5957	0.6225	0.6315
15		S-PL	0.4194	0.4278	0.4927	0.5102	0.6070	0.6092	0.6431	0.6430
16		ACL-MER	**0.5315**	**0.5549**	**0.5815**	**0.6045**	**0.6328**	**0.6382**	**0.6508**	**0.6585**
17	MMIM	No PL	0.3744	0.4199	0.4643	0.4703	0.5411	0.5443	0.6059	0.6080
18		T-PL	0.5107	0.5383	0.5184	0.5464	0.5601	0.5851	0.5958	0.6158
19		S-PL	0.3954	0.4132	0.4501	0.4571	0.5659	0.5731	0.6093	0.6096
20		ACL-MER	**0.5120**	**0.5424**	**0.5342**	**0.5579**	**0.6050**	**0.6179**	**0.6352**	**0.6408**
21	LMF	No PL	0.2744	0.3648	0.4174	0.4412	0.5360	0.5413	0.5956	0.5974
22		T-PL	0.4034	0.4420	0.5236	0.5454	0.5491	0.5777	0.5968	0.6092
23		S-PL	0.4111	0.4218	0.4828	0.4748	0.5359	0.5392	0.6050	0.6017
24		ACL-MER	**0.4118**	**0.4753**	**0.5431**	**0.5508**	**0.5852**	**0.5928**	**0.6226**	**0.6242**

Table 3: Comparison of different baselines on the MER2023 and CHERMA datasets with varying labeled data sizes. **Base Model** refers to the architecture of the student model used to perform the MER task. **Method** refers to the different Pseudo-Labeling Strategies employed. **No PL**:No Pseudo-Labeling. **T-PL**:Teacher-only Pseudo-Labeling. **S-PL**:Student-only Pseudo-Labeling. **Acc**: Accuracy.

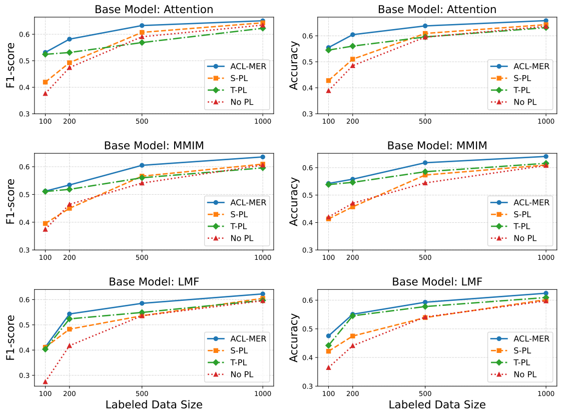


Figure 3: Impact of labeled data size on method performance on the CHERMA dataset.

Limitations of T-PL: As shown in Table 3, T-PL offers a clear advantage over both No PL and S-PL in extremely low-resource settings (e.g., lines 2 vs. 1, 3 when $N_{labeled}$=100 or 200). This indicates that even the direct, confidence-based outputs from MLLMs can provide a valuable initial supervisory signal when labeled data is extremely scarce. However, T-PL's performance advantage diminishes significantly with more labeled data, aligning with our premise that directly using MLLM outputs as hard pseudo-labels has inherent limitations. In contrast, ACL-MER's adaptive refinement strategy more effectively leverages MLLM knowledge by refining the student model's output probabilities.

Limitations of S-PL: S-PL generally performs better than No PL, demonstrating the value of conventional pseudo-labeling. However, it often struggles in the most extreme low-resource settings ($N_{labeled}$=100), where the initial student model is weak, leading to error propagation. For instance(line 7 vs 8), S-PL with MMIM on MER2023 ($N_{labeled}$=100) yields a significantly lower F1-score (0.3389) compared to ACL-MER (0.5713). ACL-MER's strength lies in its ability to refine the student's predictions with high-confidence MLLM outputs, thereby generating more reliable pseudo-labels from the outset.

Impact of Labeled Data Size Figure 3 visually illustrates the performance of ACL-MER compared to other baselines across varying labeled data sizes. Notably, the performance gap between ACL-

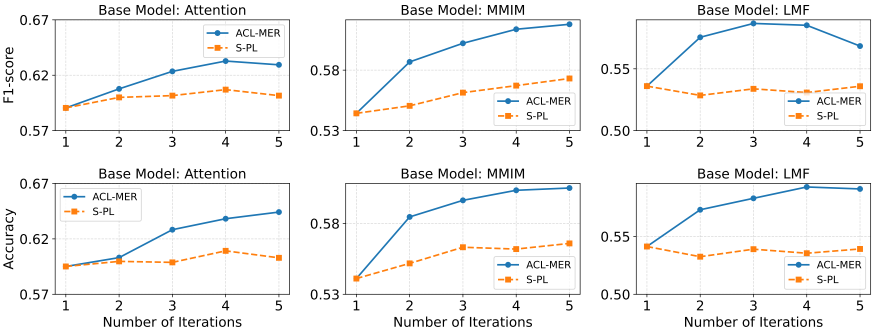


Figure 4: Performance evolution over iterations on the CHERMA dataset with $N_{labeled}$=500.

MER and other baselines is most pronounced in extremely low-resource settings ($N_{labeled}$=100 or 200). This indicates that ACL-MER's adaptive refinement strategy, which leverages MLLM knowledge and iteratively guides the student model, is particularly effective when high-quality labeled data is scarce, mitigating the challenges of noisy pseudo-labels in such environments.

Performance across Different Base Models ACL-MER consistently boosted the performance of all tested base models (Attention, MMIM, and LMF). This highlights our framework's adaptability and compatibility with various multimodal fusion architectures. While the absolute performance differed among base models, ACL-MER consistently delivered significant improvements over baselines, proving its general utility.

4.3 Ablation Studies

To comprehensively understand the contribution of each key component within ACL-MER, we conducted several ablation studies.

4.3.1 Impact of Iterative Refinement

To illustrate the effectiveness of our iterative adaptive refinement strategy, we analyze the performance of ACL-MER and S-PL over successive training iterations on the CHERMA dataset with $N_{labeled}$=500. As depicted in Figure 4, the F1-score and Accuracy of both methods are plotted against the number of iterations.

Figure 4 reveals several key insights. Firstly, ACL-MER consistently exhibits a more stable and monotonic increase in performance (both F1-score and Accuracy) across iterations compared to S-PL. While S-PL shows initial gains, its performance curve often fluctuates more, particularly in early iterations. This suggests that S-PL's pseudo-labels are less reliable and more prone to error propagation, a common challenge when relying solely on a small student model for self-supervision.

Secondly, ACL-MER achieves significantly higher final performance after a few iterations. For instance, with the Attention base model, ACL-MER reaches a considerably higher F1-score and Accuracy by the 5th iteration compared to S-PL. This demonstrates that our adaptive refinement strategy, by selectively distilling knowledge from high-confidence MLLM predictions and collaboratively guiding the student model, generates more robust and accurate pseudo-labels throughout the iterative training process. This leads to a more effective and stable learning process for the student model, enabling it to converge to superior performance by effectively leveraging the powerful, yet confidence-dependent, insights from MLLMs without succumbing to the noise prevalent in unrefined pseudo-labeling.

4.3.2 Contribution of Teacher Model Zoo

To evaluate the benefits of orchestrating a diverse teacher model zoo, we conducted an ablation study. This study compared the full ACL-MER framework (which leverages audio, visual, and textual MLLMs) against variants that utilize only a single modality-specific MLLM as a teacher. For comprehensive comparison, we also included a variant that uses Qwen2.5-Omni-3B(Jin Xu, 2025) as a single omni-modal MLLM. Further details are provided in the Appendix E. This experiment aims to demonstrate whether the collective intelligence from specialized MLLMs provides a superior supervisory signal compared to relying on a singular source. The results on the CHERMA dataset across varying labeled data sizes are presented in Table 4.

The results generally indicate that ACL-MER (Full), which incorporates a diverse teacher model

No.	Base Model	Method	$\mathrm{N_{labeled}}$=100		$\mathrm{N_{labeled}}$=200		$\mathrm{N_{labeled}}$=500		$\mathrm{N_{labeled}}$=1000	
			F1-score	Acc	F1-score	Acc	F1-score	Acc	F1-score	Acc
1	Attention	ACL-MER (Full)	**0.5315**	**0.5549**	**0.5815**	**0.6045**	**0.6328**	**0.6382**	**0.6508**	**0.6585**
2		Audio-only T.	0.4435	0.4912	0.5791	0.5974	0.5910	0.6029	0.6429	0.6460
3		Vision-only T.	0.4994	0.5519	0.5673	0.5830	0.6231	0.6336	0.6535	0.6543
4		Text-only T.	0.4794	0.5327	0.5568	0.5770	0.6221	0.6361	0.6527	0.6543
5		Omni-modal T.	0.4894	0.5208	0.5709	0.5717	0.6216	0.6267	0.6504	0.6539
6	MMIM	ACL-MER (Full)	**0.5120**	**0.5424**	**0.5342**	**0.5579**	**0.6050**	**0.6179**	**0.6352**	**0.6408**
7		Audio-only T.	0.3817	0.4448	0.4535	0.4797	0.5488	0.5577	0.6109	0.6105
8		Vision-only T.	0.4408	0.4991	0.5063	0.5323	0.5869	0.5918	0.6262	0.6292
9		Text-only T.	0.4509	0.5049	0.5192	0.5425	0.5862	0.5981	0.6262	0.6292
10		Omni-modal T.	0.4522	0.4951	0.5232	0.5397	0.5992	0.6048	0.6233	0.6267
11	LMF	ACL-MER (Full)	**0.4118**	0.4753	**0.5431**	**0.5508**	**0.5852**	**0.5928**	**0.6226**	**0.6242**
12		Audio-only T.	0.3245	0.4331	0.4098	0.4783	0.5588	0.5683	0.5979	0.6029
13		Vision-only T.	0.4022	**0.4764**	0.4737	0.5150	0.5799	0.5858	0.6222	0.6262
14		Text-only T.	0.3916	0.4356	0.4519	0.5180	0.5740	0.5817	0.6103	0.6107
15		Omni-modal T.	0.3317	0.4223	0.4529	0.5009	0.5850	0.5916	0.6148	0.6209

Table 4: Performance contribution of the teacher model zoo on CHERMA dataset. **T.** indicates Teacher.

zoo, largely achieves superior performance in most situations. This underscores the crucial role of collaborative knowledge distillation from multiple, specialized MLLMs.

Superiority of Multi-modal Teacher Ensemble: ACL-MER (Full) consistently outperforms all single-teacher variants (Audio-only T., Vision-only T., Text-only T., and Omni-modal T.). For instance, with $N_{labeled}$=100 and the Attention base model(line 1-5), ACL-MER (Full) achieves an F1-score of 0.5315, significantly higher than 0.4435 (Audio-only T.), 0.4994 (Vision-only T.), 0.4794 (Text-only T.), and 0.4894 (Omni-modal T.). Similar improvements are observed across all data sizes and base models. This suggests that different modality-specific MLLMs capture complementary aspects of emotional expression, and their combined insights lead to more robust and accurate pseudo-labels.

Limitations of Single-Modality Teachers: While individual modality-specific teachers can offer valuable supervision (often outperforming "No PL" and sometimes "S-PL" from Table 3), their performance is limited by their singular focus. For example, relying solely on audio-only or text-only teachers, especially in extremely low-resource settings, can lead to sub-optimal performance, as these teachers might miss crucial cues present in other modalities.

Limitations of Omni-modal Teacher: The "Omni-modal T." variant, which uses a single MLLM designed for general multimodal understanding, also underperforms the full ACL-MER. While omni-modal MLLMs can process audio, visual, and textual inputs simultaneously, this often comes with a high computational cost. To manage this, we had to opt for smaller model sizes for processing data, which can inherently degrade performance. Furthermore, a single omni-modal model might not possess the specialized expertise to fully capture the nuanced information critical for emotion recognition as effectively as dedicated modality-specific models. Therefore, the aggregation of high-confidence predictions from distinct, expert MLLMs proves more effective than a single, generalist MLLM in our framework.

In conclusion, this ablation study strongly confirms that the adaptive collaborative labeling strategy, by leveraging a diverse teacher model zoo of modality-specific MLLMs, is fundamental to ACL-MER's superior performance. The collective intelligence derived from these varied sources provides a richer and more reliable supervisory signal, which is particularly vital for robust learning in challenging low-resource MER scenarios.

5 Conclusion

We introduce ACL-MER, a novel framework addressing the challenge of low-resource MER. By leveraging the strengths of MLLMs within a teacher-student paradigm and introducing an adaptive collaborative refinement mechanism, ACL-MER effectively generates robust pseudo-labels, even with extremely limited labeled data. Our results on MER2023 and CHERMA demonstrate the significant advantage of this approach. This work opens exciting avenues for future research, particularly in exploring more sophisticated collaboration strategies between MLLMs and investigating the potential of this framework for other low-resource multimodal tasks.

6 Limitations

Firstly, although MLLMs offer powerful perceptual and reasoning capabilities, the reliability and discriminative power of their confidence scores are crucial for effective knowledge distillation. If teacher MLLMs are either consistently overconfident in their incorrect predictions or exhibit insufficient variance in their confidence scores to adequately differentiate between high- and low-quality outputs, our adaptive refinement strategy could be less effective. Recent work has actively focused on teaching models to express their degree of confidence in their responses, thereby enhancing the reliability and mitigating hallucinations in LLMs and MLLMs(Cheng et al., 2024a; Mahaut et al., 2024; Huang et al., 2025). Future work will investigate advanced MLLM robust confidence estimation methods to enhance the teacher's reliability and further refine the pseudo-labeling process.

Secondly, while the MLLM inference is performed only once during the initial iteration, the inherent computational cost of operating large MLLMs remains a factor. While we opted for smaller MLLM variants where necessary, deploying and inferring with multiple large models, demands substantial computational resources. This could pose a practical challenge for researchers or practitioners with limited access to high-performance computing infrastructure.

Ethical Considerations The datasets utilized in this study are exclusively for academic research in multimodal emotion recognition. Our work does not introduce direct adverse societal impacts beyond the inherent considerations of the broader field.

Acknowledgements

We thank anonymous reviewers for helpful suggestions. This work is supported by a grant from the General Office of National Language Commission Research Planning Committee (grant number: ZDA145-18).

References

Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6912–6920.

Haifeng Chen, Chujia Guo, Yan Li, Peng Zhang, and Dongmei Jiang. 2023. Semi-supervised multimodal emotion recognition with class-balanced pseudo-labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9556–9560.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024a. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024b. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Patrick Kage, Jay C Rothenberger, Pavlos Andreadis, and Dimitrios I Diochnos. 2024. A review of pseudo-labeling for computer vision. *arXiv preprint arXiv:2408.07221*.

Sun-Mee Kang and Phillip R Shaver. 2004. Individual differences in emotional complexity: Their psychological implications. *Journal of personality*, 72(4):687–726.

H Toprak Kesgin and M Fatih Amasyali. 2022. Investigating semi-supervised learning algorithms in text datasets. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal sentiment analysis: A survey. *Displays*, 80:102563.

Dong-Hyun Lee and 1 others. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, and 1 others. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, and 1 others. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614.

Zheng Lian, Haiyang Sun, Licai Sun, Lan Chen, Haoyu Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, and 1 others. 2024a. Open-vocabulary multimodal emotion recognition: Dataset, metric, and benchmark. *arXiv preprint arXiv:2410.01495*.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2024b. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*.

Jingjun Liang, Ruichen Li, and Qin Jin. 2020. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2852–2861.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. *arXiv preprint arXiv:2406.13415*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

David MW Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 658–670, Toronto, Canada. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024b. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2282–2291.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Sheng Wu, Xiaobao Wang, Longbiao Wang, Dongxiao He, and Jianwu Dang. 2024. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. *arXiv preprint arXiv:2412.10460*.

Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Zongyang Ma, Wanxiang Che, and Bing Qin. 2024. Large language models meet text-centric multimodal sentiment analysis: A survey. *arXiv preprint arXiv:2406.08068*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. 2022. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. 2021. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459.

A Prompt

A.1 Prompt for Audio LLMs

~~~
As an expert in the field of emotion, please focus on the audio information within the video to analyze clues related to personal emotions. Based on your analysis, please provide the following:

Detailed Description: Please analyze the speaker's emotional state in detail, step by step. First, describe the observable sound characteristics in the audio (e.g., pitch, volume, speech rate, pauses); then infer the emotional states or psychological dynamics that these characteristics might indicate; finally, combine the above analysis to provide a comprehensive emotional description. Please ensure logical clarity and coherence in the analysis process.

Predicted Emotional State: Select the most likely emotion from the following emotional states: ['neutral', 'happy', 'angry', 'sad', 'disgust', 'fear', 'surprise'].

Prediction Confidence: Based on your analysis, estimate the confidence level of the prediction, using a score from 0 to 5, where 0 indicates no confidence and 5 indicates complete confidence.

Please strictly adhere to the following output format:
```json
{
"description": "[Please provide your analysis here]",
"emotion": "[Select an emotion from the candidate set]",
"confidence": [0-5]
}
```
~~~

A.2 Prompt for Vision LLMs

~~~~
As an expert in the field of emotion, please focus on information from the video such as facial expressions, body language, and environmental cues to analyze clues related to personal emotions. Based on your analysis, please provide the following:

Detailed Description: Please focus on facial expressions, body movements, environmental information, and other visual cues in the video, analyzing clues related to personal emotions step-by-step. First, describe the observable visual features in the video; then infer the emotional states or psychological dynamics that these features might indicate; finally, combine the above analysis to provide a comprehensive emotional description. Please ensure logical clarity and coherence in the analysis process.

Predicted Emotional State: Select the most likely emotion from the following emotional states: ['neutral', 'happy', 'angry', 'sad', 'disgust', 'fear', 'surprise'].

Prediction Confidence: Based on your analysis, estimate the confidence level of the prediction, using a score from 0 to 5, where 0 indicates no confidence and 5 indicates complete confidence.

Please strictly adhere to the following output format:
```json
{
"description": "[Please provide your analysis here]",
"emotion": "[Select an emotion from the candidate set]",
"confidence": [0-5]
}
```
~~~~

A.3 Prompt for LLMs

~~~~
As an expert in the field of emotion, please analyze clues related to an individual's emotional state based on the provided information.

The video description is as follows:<video caption>
The audio description is as follows: <audio caption>
The speaker's lines are as follows: <subtitle>

Please note that the descriptions of the video and audio may be inaccurate, or there may be no descriptions provided for them. Based on your analysis, please provide the following content:

Detailed Description: Please analyze the speaker's emotional state in detail, step by step. First, analyze the content of the speaker's dialogue; then infer the emotional states or psychological dynamics that these features might indicate; finally, combine the above analysis to provide a comprehensive emotional description. Please ensure logical clarity and coherence in the analysis process.

Predicted Emotional State: Select the most likely emotion from the following emotional states: ['neutral', 'happy', 'angry', 'sad', 'disgust', 'fear', 'surprise'].

Prediction Confidence: Based on your analysis, estimate the confidence level of the prediction, using a score from 0 to 5, where 0 indicates no confidence and 5 indicates complete confidence.

Please strictly adhere to the following output format:
```json
{
"description": "[Please provide your analysis here]",
"emotion": "[Select an emotion from the candidate set]",
"confidence": [0-5]
}
```
~~~~

Here, `<subtitle>`refers to the spoken words of the speaker, `<video caption>` is the description generated by prompting Visual LLMs, and `<audio caption>` is the description generated by prompting Audio LLMs.

B Preliminary Experiments

To motivate our proposed ACL-MER framework and highlight the challenges it aims to address, we conducted two sets of preliminary experiments.

Preliminary Experiments in Figure 1 We first investigated the performance of a conventional MER model under varying amounts of labeled training data. The experimental setup for these preliminary tests was identical to that described for the main experiments in Section 4.1. Specifically, for the base model, we utilized the attention-based model to perform the MER task. As depicted in Figure 1, both Accuracy and F1-score exhibit a substantial drop when only a few hundred labeled samples are available, indicating the limitations of pseudo-labeling in low-resource scenarios. This highlights a core limitation of conventional pseudo-labeling: it struggles to produce reliable supervision when labeled data is extremely scarce.

Preliminary Experiments in Table 1 Next, we explored the capabilities of modality-specialized MLLMs in zero-shot emotion recognition. The experimental setup for these preliminary tests was consistent with the MLLMs used as teachers in our main experiments (detailed in Section 4.1). The evaluation was performed on the MER2023 test dataset. We prompted each MLLM to generate emotion predictions along with scalar confidence scores. Subsequently, we analyzed the Accuracy of these MLLM predictions across different confidence levels. As shown in Table 1, high-confidence outputs (e.g., confidence = 5) achieved accuracies above 0.77, outperforming small models trained on 500 labeled samples. However, these predictions only covered a limited portion of the data (e.g., 11.3% for audio, 24.3% for vision), and Accuracy degraded sharply when lower-confidence outputs were included. These observations suggest that while MLLMs can offer valuable supervision, their coverage is limited and their reliability is confidence-dependent.

Significance of Preliminary Findings Through these two preliminary experiments, we observed the distinct advantages and inherent limitations of both traditional smaller MER models and state-of-the-art multimodal large language models in low-resource settings. Specifically, small models struggle with scarce labeled data, while MLLMs provide high-quality but limited coverage and confidence-dependent predictions. These observations provided the critical motivation and foundational insights for the development of our proposed ACL-MER framework, which aims to synergistically combine the strengths of both paradigms to overcome their individual shortcomings.

C Base Models

This section details the base models employed in experiments.

- **Attention** (Vaswani, 2017): The Attention model combines information from different modalities by first processing each separately with an MLP encoder. These processed features are then joined together, and an attention mechanism learns to weigh their importance.

- **MultiModal InfoMax (MMIM)** (Han et al., 2021): MMIM enhances multimodal sentiment analysis by hierarchically maximizing mutual information. This maximization occurs both between unimodal inputs and between the multimodal fusion results and the original unimodal inputs.

- **Low-rank Multimodal Fusion(LMF)** (Liu et al., 2018): LMF is an efficient multimodal fusion method that leverages low-rank tensor decomposition to integrate heterogeneous modalities while significantly reducing computational complexity.

D Implementation Details

D.1 Environment

We used Python 3.8 with PyTorch 1.13.

D.2 Hardware Configurations

All experiments are conducted using a Tesla V100 GPU.

D.3 Hyperparameter Details

Model training was performed using the Adam optimizer(Kingma and Ba, 2014) with a learning rate of 1×10^{-4} and a batch size of 32. A weight decay (L2 regularization) of 1×10^{-4} was applied to the optimizer. Training proceeded for 50 epochs in each semi-supervised learning iteration.

The ACL-MER framework operates over a maximum of 5 semi-supervised iterations. Within each iteration, the probability adjustment value $p_{init} = 0.6$ and decays linearly by a rate of $\gamma = 0.1$ per iteration, with a lower bound of $p_{min} = 0.2$.

For pseudo-label selection, a selection threshold θ was used to identify high-confidence samples. For $N_{labeled}$=100, θ was set to 0.8; otherwise, $\theta = 0.995$. To mitigate class imbalance in the pseudo-labeled dataset, up to $top\_k = 200$ samples were selected per class, based on their highest refined probabilities.

For the inference phase of the teacher MLLMs, the following hyperparameters were set: temperature was 0.6 and top_p was 0.95. To enhance the reliability of MLLM predictions, we employed a self-consistency method(Wang et al., 2022): each teacher model performed inference three times for every sample. The average confidence score was computed only when all three predicted emotion labels were consistent. A confidence threshold of $c_t = 4$ was applied, meaning only predictions with a confidence score greater than 4 were considered reliable enough to be incorporated into the ACL-MER refinement process.

To ensure the robustness and reliability of our results, we employed a 5-fold cross-validation strategy. This involved partitioning the training dataset into five equally sized folds. In each of the five cross-validation runs, one fold was held out as the validation set, and the remaining four folds were used for training. This process was repeated five times, with each fold serving as the validation set exactly once. When presenting our results, we provide mean performance across the 5 cross-validation folds.

All datasets and pre-trained models utilized in this study were accessed and used in accordance with their specified research-only licenses and intended academic purposes.

D.4 Data and Code Availability

The code for our ACL-MER framework, including training scripts and model configurations, will be made publicly available on GitHub upon acceptance. All released code will be licensed under the MIT License. The datasets used in this study, MER2023 and CHERMA, are publicly available under their respective licenses, and our usage adheres to their terms.

E Details for Omni-model Teacher in Ablation Studies

E.1 Model Selection

Table 5 presents the minimum GPU memory requirements for Qwen-Omni models. Given our computational resource constraints and the extended durations of videos within our dataset, we selected Qwen2.5-Omni-3B as our Omni-model Teacher, rather than the larger 7B version.

Model	Precision	15(s) Video	30(s) Video
Qwen-Omni-3B	FP32	89.10 GB	Not Recommend
Qwen-Omni-3B	BF16	18.38 GB	22.43 GB
Qwen-Omni-7B	FP32	93.56 GB	Not Recommend
Qwen-Omni-7B	BF16	31.11 GB	41.85 GB

Table 5: Minimum GPU memory requirements for Qwen-Omni models(Jin Xu, 2025).

E.2 Inference Parameters

The inference parameters were set as follows: temperature = 0.6, top_p = 0.95.

E.3 Prompt

As an expert in the field of emotion, please focus on the visuals and audio within the video, as well as the speaker's dialogue, to analyze clues related to personal emotions. Based on your analysis, please provide the following:

Detailed Description: Please analyze the speaker's emotional state in detail, step by step. First, describe the observable sound characteristics in the video (e.g., pitch, volume, speech rate, pauses) and visual characteristics (e.g., facial expressions, body movements, environmental information); then infer the emotional states or psychological dynamics that these characteristics might indicate; finally, combine the above analysis to provide a comprehensive emotional description. Please ensure logical clarity and coherence in the analysis process.

Predicted Emotional State: Select the most likely emotion from the following emotional states: ['neutral', 'happy', 'angry', 'sad', 'disgust', 'fear', 'surprise'].

Prediction Confidence: Based on your analysis, estimate the confidence level of the prediction, using a score from 0 to 5, where 0 indicates no confidence and 5 indicates complete confidence.

Please strictly adhere to the following output format:

~~~
```json
{
"description": "[Please provide your analysis here]",
"emotion": "[Select an emotion from the candidate set]",
"confidence": [0-5]
}
```
~~~

F Additional Main Experimental Results on the MELD dataset

We have conducted additional experiments on the widely used English conversational emotion recognition dataset, MELD(Poria et al., 2019). The experimental setup remains identical to that used for the other datasets.

As shown in Table 6, our proposed ACL-MER method consistently outperforms all baselines across various labeled data sizes on MELD. This result further demonstrates the robustness and generalization capability of ACL-MER beyond the Chinese datasets (MER2023 and CHERMA).

G Additional Baseline Results

Table 3 primarily compares different pseudo-labeling methods; that is, methods based on the self-training paradigm. Here, we supplement those results by also including the co-training paradigm, following the work of (Kesgin and Amasyali, 2022).

Co-training works by training two view-specific classifiers; each classifier then confidently labels unlabeled samples for the other, and they re-train alternately to expand the labeled set. In our implementation, we created two different feature views: one model was trained on a single modality, and the second was trained on the remaining two modalities. The final performance was evaluated using model fusion by averaging the output probabilities of the two models.

As the results in Table 7 show, our proposed ACL-MER method consistently outperforms the

No.	Base Model	Method	$N_{labeled}$=100		$N_{labeled}$=200		$N_{labeled}$=500		$N_{labeled}$=1000	
			F1-score	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**
1	Attention	No PL	0.4398	0.4500	0.4646	0.4879	0.5034	0.5301	0.5131	0.5431
2		T-PL	0.3728	0.3453	0.3853	0.3756	0.4636	0.4626	0.5125	0.5228
3		S-PL	0.4511	0.4588	0.4797	0.5094	0.5101	0.5339	0.5230	0.5554
4		ACL-MER	**0.4733**	**0.4872**	**0.4822**	**0.5117**	**0.5246**	**0.5489**	**0.5400**	**0.5738**

Table 6: Comparison of different baselines on the MELD datasets with varying labeled data sizes.

No.	Method		N_{train}=100		N_{train}=200		N_{train}=500		N_{train}=1000	
			F1-score	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**
1	self-training		0.4194	0.4278	0.4927	0.5102	0.6070	0.6092	0.6431	0.6430
2		v | ta	0.5197	0.4599	0.5369	0.4675	0.6161	0.5858	0.6494	0.6400
3	co-training	a | tv	0.5299	0.4799	0.5727	0.4788	0.6052	0.5750	0.6407	0.6324
4		t | av	0.5155	0.4377	0.5522	0.4816	0.5818	0.5508	0.6303	0.6250
5	ACL-MER		**0.5315**	**0.5549**	**0.5815**	**0.6045**	**0.6328**	**0.6382**	**0.6508**	**0.6585**

Table 7: Comparison of different baselines on the CHERMA datasets with varying labeled data sizes. Attention model is used as student model. In the co-training methods, "a", "v", and "t" denote audio, visual, and text modalities, respectively. For example, "a | tv" means one model uses audio as its view, while the other uses a combination of text and visual modalities.

self-training and co-training baselines across all experimental settings. We believe this is due to the inherent information loss in co-training, as each model only utilizes a subset of the available modalities. This is particularly evident when the text modality is used as a single-view input, which leads to weaker performance and limits the overall effectiveness of the approach.

H Impact of Different Prompt Designs

In our original ACL-MER approach, we adopted self-consistency and CoT reasoning to enhance the quality of emotion generation by MLLMs. To demonstrate their effectiveness, we have now included experiments evaluating the performance without self-consistency and without CoT.

The results of this ablation study are presented in Table 8. The findings clearly indicate that both the Self-Consistency mechanism and CoT are crucial for achieving the best performance across all training data sizes, consistently showing the highest F1-scores and Accuracy when both are employed in the full ACL-MER method.

I Hyperparameter Sensitivity Analysis

To assess the robustness and stability of our proposed ACL-MER framework, we conduct a comprehensive hyperparameter sensitivity analysis. We systematically vary one hyperparameter at a time while keeping all others at their empirically determined default values ($p_{min} = 0.2$, $p_{init} = 0.6$, $\gamma = 0.1$, $\theta = 0.995$, and $top\_k = 200$), consistent with our main experiments. All analyses are performed on the CHERMA dataset with $N_{labeled} = 200$ using the Attention base model, where the low-resource setting makes the impact of these parameters most pronounced. The results are presented in Figure 5 .

Impact of Pseudo-Label Selection Threshold (θ) The pseudo-label selection threshold, θ, is a critical hyperparameter that mediates the trade-off between the **quality** (high confidence) and **quantity** (coverage) of pseudo-labels used to enhance the training data. Figure 5(a) illustrates the effect. The analysis of ACL-MER's performance across the extended θ range, from 0.8 to 1.2, reveals two distinct behaviors. In the conventional high-confidence range ($\theta \in [0.8, 0.995]$), ACL-MER demonstrates a remarkable stability in performance, with both F1 scores and Accuracy remaining robustly high. For instance, the F1 score only minimally fluctuates between 0.5867 ($\theta = 0.8$) and 0.5815 ($\theta = 0.995$). Crucially, the optimal performance is pinpointed at $\theta = 0.995$ (Accuracy: 0.6045), confirming that maintaining a sufficiently high quality standard for pseudo-labels yields the best results without unduly sacrificing data coverage. However, the performance undergoes a significant degradation as θ is pushed into the extreme confidence range ($\theta > 0.995$). When θ exceeds 1.0, the threshold becomes so restrictive that only pseudo-labels with extremely high confidence—primarily those contributed by the more conservative MLLMs—are accepted. This ultra-high confidence criterion leads to

No.	Methods	N$_{\text{train}}$=100		N$_{\text{train}}$=200		N$_{\text{train}}$=500		N$_{\text{train}}$=1000	
		F1-score	Acc	F1-score	Acc	F1-score	Acc	F1-score	Acc
1	ACL-MER	**0.5315**	**0.5549**	**0.5815**	**0.6045**	**0.6328**	**0.6382**	**0.6508**	**0.6585**
2	without self-consistency	0.4933	0.5358	0.5694	0.5941	0.6093	0.6340	0.6417	0.6548
3	without CoT	0.5307	0.5528	0.5850	0.5815	0.6312	0.6333	0.6496	0.6500

Table 8: Ablation study results on different prompt designs on CHERMA dataset. Attention model is used as student model.

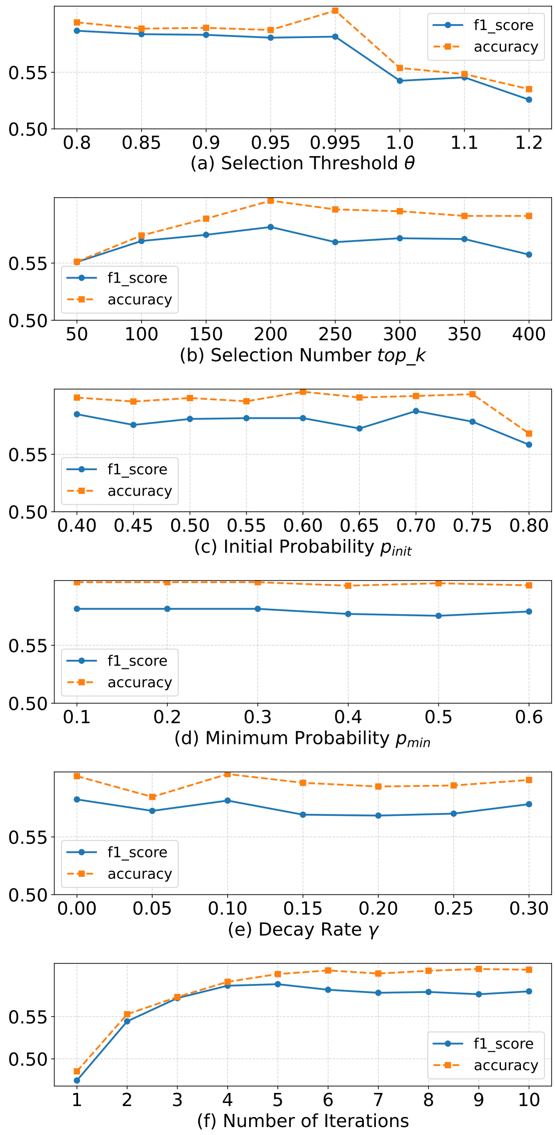


Figure 5: Hyperparameter sensitivity analysis of ACL-MER on the CHERMA dataset with $N_{labeled} = 200$. (a) Select Threshold θ, (b) Select Number top_k, (c) Initial Probability p_{init}, (d) Minimum Probability p_{min}, (e) Decay Rate γ, and (f) Number of Iterations.

a pronounced scarcity of reliable samples, thereby sharply diminishing the contribution of the student model and the overall effectiveness of the collaborative learning process. The F1 score drops notably from 0.5815 ($\theta = 0.995$) to 0.5425 ($\theta = 1.0$) and continues to fall to 0.5258 ($\theta = 1.2$). This decline is attributed to the excessively high θ values causing the iterative self-training procedure to stall prematurely due to an insufficient introduction of new, high-quality pseudo-labels, ultimately impairing the final model performance. Therefore, while ACL-MER is robust within a wide high-confidence band, setting the threshold too high detrimentally starves the model of necessary training signals.

Impact of Pseudo-Label Selection Number (top_k) As shown in Figure 5(b), the number of pseudo-labels selected per class (top_k) significantly influences performance. Performance gradually improves as top_k increases from 50, reaching its peak at approximately $top\_k = 200$, then showing a slight decrease. Selecting too few samples (e.g., $top\_k = 50$) limits the amount of new information available for the student model, while selecting too many (e.g., $top\_k = 400$) may introduce more noisy pseudo-labels, particularly in imbalanced datasets, despite the class-balanced selection strategy. The optimal range indicates that a moderate number of high-confidence samples per class is most beneficial.

Impact of Initial Probability Adjustment (p_{init}) Figure 5(c) demonstrates the sensitivity to the initial probability adjustment value, p_{init}. Performance shows a clear peak when p_{init} is around 0.6 to 0.65. A lower p_{init} (e.g., 0.4) might mean that the MLLM's strong guidance is not sufficiently leveraged, failing to effectively refine the student's initial predictions. Conversely, a higher p_{init} (e.g., 0.8) could overly prioritize the MLLM's predictions, potentially overriding the student's nascent learning or introducing noise if MLLM confidence is not perfectly correlated with accuracy across all samples. This highlights the importance of carefully tuning p_{init} to balance MLLM guidance with student self-learning.

Impact of Minimum Probability (p_{min}) Figure 5(d) reveals that ACL-MER is notably robust to changes in the minimum probability adjustment, p_{min}. Across the tested range from 0.1 to 0.6, both F1-score and Accuracy remain remarkably stable. This observed robustness can be attributed to p_{min} primarily influencing the probability adjustments in the later iterations of the training process. In the initial phases, the probability adjustment p_t is largely determined by p_{init} and the decay rate γ. Once p_t decays to the floor set by p_{min}, the impact of the teacher models becomes less pronounced. Since significant learning and pseudo-label generation occur in earlier iterations with higher adjustment values, varying this lower bound in the later stages has a minimal effect on the overall final performance within this tested range.

Impact of Decay Rate (γ) We investigated the influence of the decay rate γ. As depicted in Figure (e), the model's performance, as measured by both F1-score and Accuracy, demonstrates a relatively stable trend across the tested range of γ values. While there are minor fluctuations, both metrics generally maintain strong performance. A notable peak in both Accuracy and F1-score is observed at a moderate decay rate of $\gamma = 0.1$. Performance remains robust around this point, suggesting that the framework is not overly sensitive to precise tuning of this hyperparameter. Based on these empirical observations, we selected $\gamma = 0.1$ for our main experiments to achieve optimal balance.

Impact of Number of Iterations As depicted in Figure 5(f), the performance of ACL-MER improves significantly in the initial iterations, with a sharp increase in both F1-score and Accuracy from iteration 1 to 3. The gains continue, albeit at a slower pace, stabilizing and reaching a plateau around 5 to 6 iterations. Beyond this point, further iterations yield marginal improvements or even a slight decline, indicating that the model has converged or is beginning to suffer from potential error propagation. This analysis confirms the effectiveness of the iterative refinement process and helps determine an appropriate stopping criterion, demonstrating that significant gains are achieved within a reasonable number of iterations.

In conclusion, our hyperparameter sensitivity analysis demonstrates that ACL-MER is relatively robust to several key parameters like θ, p_{min} and γ, while showing moderate sensitivity to p_{init}, top_k, and number of iterations. Identifying these optimal ranges is crucial for maximizing performance in low-resource MER. The iterative nature of ACL-MER also proves highly effective, yielding substantial improvements within a limited number of cycles.