Resource-Limited Joint Multimodal Sentiment Reasoning and Classification via Chain-of-Thought Enhancement and Distillation

Haonan Shangguan* , **Xiaocui Yang*†**, **Shi Feng, Daling Wang, Yifei Zhang, Feiliang Ren, Ge Yu**
School of Computer Science and Engineering, Northeastern University, Shenyang, China
20217424@stu.neu.edu.cn, yangxiaocui@cse.neu.edu.cn, fengshi@cse.neu.edu.cn
{wangdaling,zhangyifei,renfeiliang,yuge}@cse.neu.edu.cn

Abstract

Current approaches for Multimodal Sentiment Analysis (MSA) primarily leverage the knowledge and reasoning capabilities of parameter-heavy (Multimodal) LLMs for classification, overlooking autonomous multimodal sentiment reasoning generation in resource-constrained environments. In this paper, we focus on the Resource-Limited Joint Multimodal Sentiment Reasoning and Classification task, JMSRC, which simultaneously performs multimodal sentiment reasoning chain generation and sentiment classification only with a lightweight model. We propose a Multimodal Chain-of-Thought Reasoning Distillation model, MulCoT-RD, designed for JMSRC that employs a "Teacher-Assistant-Student" distillation paradigm to address deployment constraints in resource-limited environments. We first leverage a high-performance Multimodal Large Language Model (MLLM) to generate the initial reasoning dataset and train a medium-sized assistant model with a multi-task learning mechanism. A lightweight student model is jointly trained to perform efficient multimodal sentiment reasoning generation and classification. Extensive experiments on four datasets demonstrate that MulCoT-RD[1], with only 3B parameters, achieves strong performance on JMSRC while exhibiting robust generalization and enhanced interpretability.

1 Introduction

With the proliferation of social media and multimedia content, Multimodal Sentiment Analysis (MSA) has emerged as a critical research area attracting significant academic and industry attention Yang et al. (2024); Amiriparian et al. (2024). MSA of text-image pairs can be categorized into coarse-grained and fine-grained approaches based on sentiment targets. Coarse-grained MSA (Yang et al., 2021; Zhang et al., 2023) identifies the overall sentiment of text-image pairs, while fine-grained MSA, or Multimodal Aspect-Based Sentiment Classification (MASC) (Zhou et al., 2023; Wang et al., 2024; Yang et al., 2025b), analyzes sentiment toward specific aspect terms within textual content.

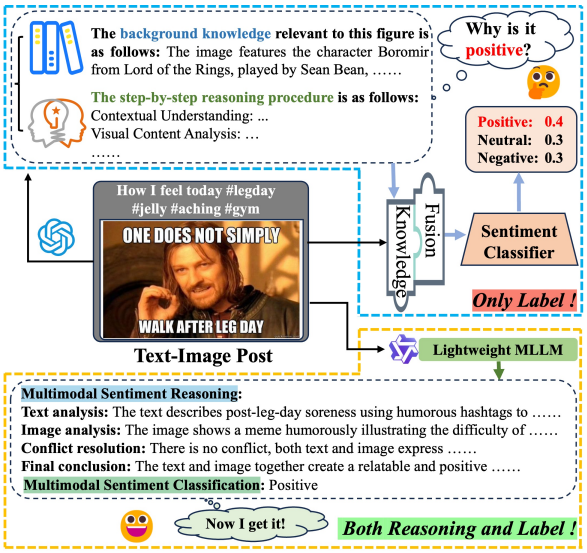

Figure 1: Leveraging reasoning (blue dashed line) vs. Generating reasoning chain (yellow dashed line).

Most existing methods enhance MSA through multimodal representation learning (Zhang et al., 2022; Manzoor et al., 2023) and fusion (Huang et al., 2020; Zhang et al., 2023), employing separate encoders to extract unimodal representations, then integrating them using fusion strategies such as gating mechanisms (Kumar and Vepa, 2020), cross-modal attention (Ju et al., 2021), and graph neural networks (Yang et al., 2021). While these approaches advance MSA performance, they face a fundamental limitation: inability to model intra-modal and cross-modal sentiment reasoning processes that explain why users experience particular sentiments. These models typically operate as "black boxes" for sentiment classification, obscuring the specific contributions of each modality

*Equal contribution.
†Corresponding author.
[1]The code and demo are available via https://github.com/123sghn/MulCoT-RD.

Findings of the Association for Computational Linguistics: ACL 2026, pages 15971–15986

and interaction mechanisms in sentiment decisions due to their lack of explicit modeling of sentiment presentation and reasoning chain across modalities.

Building upon LLMs, Multimodal Large Language Models (MLLMs) (Hurst et al., 2024; Wu et al., 2024; Bai et al., 2025) demonstrate remarkable performance across diverse multimodal tasks, including recommendation systems (Ye et al., 2025), sentiment analysis (Wang et al., 2024), and mental health assessment (Zhang et al., 2024). As shown in Figure 1 (blue box), current methods leverage high-performing MLLMs, like GPT-4o, to inject world knowledge or Chain-of-Thought (CoT) (Wei et al., 2022) reasoning into pre-trained language models for MSA improvement (Wang et al., 2024; Li et al., 2025a), yet fail to transfer superior reasoning capabilities. Existing research (Li et al., 2025b) shows that lightweight MLLMs ($\leq$3B parameters) exhibit limited CoT reasoning capabilities, necessitating reliance on models with superior reasoning abilities. However, closed-source models incur substantial costs, while large-scale MLLMs require extensive computational resources, limiting deployment in resource-constrained environments. Developing lightweight MLLMs (e.g., 3B parameters) that autonomously generate high-quality multimodal sentiment reasoning while maintaining high MSA performance represents a major challenge, as highlighted in the yellow box of Figure 1.

To address these challenges, we focus on the **Resource-Limited Joint Multimodal Sentiment Reasoning and Classification (JMSRC)** task, which simultaneously performs multimodal sentiment reasoning generation and classification using only a lightweight MLLM. We introduce the **Multimodal Chain-of-Thought Enhancement with Reasoning Distillation (MulCoT-RD)** framework for JMSRC, illustrated in Figure 2, while leveraging Reasoning Distillation (RD) with the Teacher-Assistant-Student pattern to enable lightweight MLLMs to autonomously generate high-quality sentiment reasoning (for the second challenge). The MulCoT-RD comprises two core modules. **(1) Multimodal CoT Enhancement Module**: We design a two-stage module using structured prompt templates with task decomposition, reasoning guidance, conflict mediation steps, and adaptive retry control. It guides the high-performance closed-source or large-scale open-source MLLM as a teacher model to generate logically coherent multimodal sentiment reasoning. **(2) Multimodal Sentiment Reasoning Distillation Module**: Considering teacher model limitations in providing soft labels and intermediate representations, data scarcity, and inference costs, we introduce a medium-sized open-source MLLM as an assistant model, and use it to synthesize high-quality data. Through multi-task learning, the assistant model jointly enhances sentiment label prediction accuracy and reasoning quality. For efficient deployment in resource-constrained environments, we employ joint optimization combining hard labels with soft labels from the assistant model to transfer reasoning capabilities to a lightweight student MLLM, achieving optimal balance among classification performance, interpretability, and deployment efficiency. Our contributions are summarized as follows:

- We focus on joint multimodal sentiment reasoning and classification in resource-constrained scenarios and construct a high-quality sentiment reasoning dataset.
- We propose the Multimodal Chain-of-Thought Enhancement with Reasoning Distillation, MulCoT-RD, framework for JMSRC. Multi-task learning and joint optimization improve the sentiment classification and reasoning capabilities of the model.
- Comprehensive experiments across multiple MSA datasets demonstrate that our lightweight 3B-parameter MLLM achieves superior sentiment classification performance while maintaining high interpretability.

2 Related Work

2.1 Multimodal Sentiment Analysis

The MSA development can be broadly divided into two stages: the era of pre-trained language models (PLMs) and the era of large language models (LLMs). During the PLMs era, MSA methods typically utilize a dedicated encoder for each modality to extract representations, with a primary focus on multimodal fusion and cross-modal alignment. (Zhang et al., 2023; Xiao et al., 2023; Zhou et al., 2023). The emergence of LLMs has opened new possibilities for MSA. However, existing methods typically rely on MLLMs to generate valuable knowledge (Wang et al., 2024) or reasoning (Pang et al., 2024; Li et al., 2025a), which is then injected into pre-trained language models to improve MSA, rather than enabling autonomous sentiment reasoning. This results in limited interpretability. To

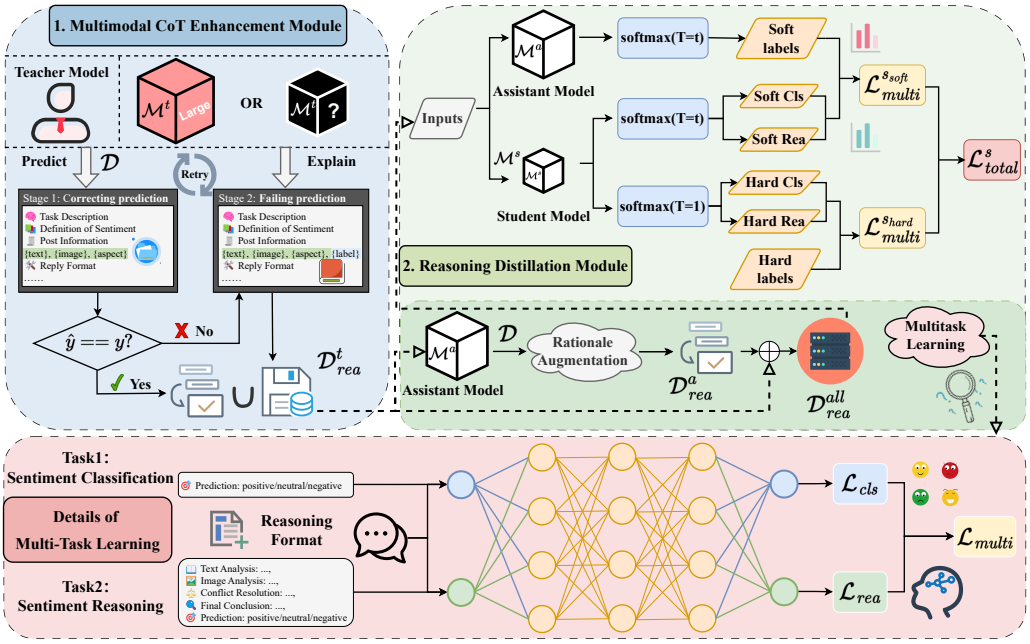

Figure 2: MulCoT-RD comprises two core modules, i.e., (1) Multimodal CoT Enhancement Module, (2) Reasoning Distillation Module (Assistant Model with Multi-Task Learning, Student Model with Joint Learning).

our knowledge, Emotion-LLaMA is the first LLM-based model for multimodal emotion recognition and explanation, but requires modality-specific representation learning, pre-training, and instruction tuning (Cheng et al., 2024). Models with superior reasoning capabilities are often computationally expensive or have large parameter counts that complicate deployment. We focus on using lightweight MLLM to simultaneously achieve efficient and autonomous generation of high-quality multimodal sentiment reasoning and classification.

2.2 Reasoning Distillation

Knowledge Distillation (KD) (Hinton et al., 2015) has proven effective for compressing language models by transferring predictive behaviors, such as soft labels or hidden representations, from larger teacher models to smaller student models. Current KD techniques for PLMs focus on distilling soft labels (Sanh et al., 2019; Gu et al., 2023) or representations (Wang et al., 2020b,a; Kim et al., 2022), but require access to the teacher model's internal parameters. This dependency creates significant challenges when applying KD to closed-source LLMs. Reasoning distillation offers an alternative approach, enabling smaller student models to acquire reasoning capabilities by fine-tuning on reasoning processes from a teacher model instead of relying on soft labels (Magister et al., 2022; Li et al., 2023; Lee et al., 2024; Chenglin et al., 2024). In our work, we leverage an intermediate-sized model with multi-task learning as an assistant to both supplement soft-label distillation signals from the teacher model and generate higher-quality data to address reasoning data scarcity.

3 Method

To achieve an effective integration of task performance, interpretability, and deployment efficiency, we introduce the Multimodal Chain-of-Thought Enhancement with Reasoning Distillation (MulCoT-RD) framework for JMSRC, as shown in Figure 2, comprising the Multimodal CoT Enhancement Module and the Reasoning Distillation Module.

3.1 Task Definition

Given a dataset $\mathcal{D} = \{(x_i, L_i)\}_{i=1}^{N}$ containing N samples, each sample x_i consists of text T_i, image I_i, aspect term $[A_i]$ (provided only in fine-grained MSA), and sentiment label L_i. The JMSRC task is formulated as follows:

$$\mathcal{M}\left(T_i,\ I_i, [A_i]\right) \Rightarrow \left(R_i, \hat{y}_i\right), \tag{1}$$

where R_i denotes the corresponding sentiment reasoning, and $\hat{y}_i$ denotes the predicted sentiment label by MLLM $\mathcal{M}$.

3.2 Multimodal CoT Enhancement

We propose a two-stage multimodal CoT enhancement module to synthesize high-quality sentiment reasoning data. The corresponding prompts are illustrated in Figure 6 provided in Appendix F. **In**

the first stage, we perform reasoning path generation in a label-free setting using a high-performance MLLM as the teacher model $\mathcal{M}^t$. We employ a structured CoT prompt template $\mathcal{T}_{pre}$ for **prediction**, comprising the basic template $\mathcal{T}_b$ (including Task Description, Sentiment Definition, and Reasoning Format) and the specific prediction prompt $\mathcal{P}_{pre}$. This template guides the model through text analysis, image analysis, conflict resolution, and conclusion generation, ensuring logically coherent and interpretable reasoning.

$$c_i^{t_1}, \hat{y}_i^t = \mathcal{M}^t\left(x_i; \mathcal{T}_{pre}\right), \tag{2}$$

where $c_i^{t_1}$ represents the CoT reasoning process generated in the first stage, and $\hat{y}_i^t$ indicates the predicted sentiment label for the i-th sample.

For correctly predicted samples, the generated reasoning paths are directly retained for subsequent training, thereby constructing the first-stage training set, $\mathcal{D}_{rea}^{t1}$.

$$\mathcal{D}_{rea}^{t_1} = \left\{\left(x_i, c_i^{t_1}, \hat{y}_i^t\right) \mid \hat{y}_i^t = L_i\right\}_{i=1}^{N_{t_1}}. \tag{3}$$

Misclassified samples often reflect complex cases with ambiguous boundaries or cross-modal conflicts, or semantic ambiguity. Guiding the model to learn causally consistent reasoning on these challenging examples can enhance its understanding and robustness in complex scenarios. Therefore, we design a second stage where, for samples with incorrect predictions, the ground truth label, L_i, is introduced and an explain template, $\mathcal{T}_{exp}$, is constructed to guide the model in generating a supervised reasoning process, $c_i^{t_2}$, conditioned on the correct label.

$$\begin{cases} c_i^{t_2}, \hat{y}_i^t = \mathcal{M}^t\left(x_i, L_i; \mathcal{T}_{exp}\right) \\ \mathcal{D}_{rea}^{t_2} = \left\{\left(x_i, c_i^{t_2}, L_i\right)\right\}_{i=1}^{N_{t_2}}, \end{cases} \tag{4}$$

where $N = N_{t1} + N_{t2}$; $\mathcal{T}_{exp}$ is constructed by the basic template, $\mathcal{T}_b$, and the specific reasoning prompt, $\mathcal{P}_{exp}$, as shown in Figure 6 provided in Appendix F.

The two-stage datasets are merged to obtain the reasoning dataset $\mathcal{D}_{rea}^t = \mathcal{D}_{rea}^{t_1} \cup \mathcal{D}_{rea}^{t_2}$. To improve sentiment reasoning and label prediction reliability, we introduce an adaptive retry controller, ARC, that automatically regenerates outputs when MLLMs produce incomplete structures or invalid labels until a valid result is obtained or the retry limit is reached, ensuring generation quality while controlling computational overhead. The details of ARC are provided in Appendix E.

3.3 Multimodal Sentiment Reasoning Distillation

Closed-source teacher models limit knowledge extraction due to restricted intermediate representations, while open-source models with strong reasoning often require large parameters (Li et al., 2025b), hindering efficient deployment. To address multimodal sentiment reasoning data scarcity and the absence of soft labels, we introduce reasoning distillation (Lee et al., 2024) to train an **assistant model with multi-task learning**, as illustrated in the middle right of Figure 2, enhancing data diversity. A **student model with joint learning**, as shown in the upper right of Figure 2, adapts to resource-constrained environments while inheriting the assistant model's sentiment reasoning and classification capabilities.

3.3.1 Assistant Model with Multi-Task Learning

We propose a multi-task learning framework that shares hard parameters to train the assistant model, $\mathcal{M}^a$, for JMSRC that jointly optimizes two complementary tasks, including multimodal sentiment reasoning and classification, as shown in the lower part of Figure 2.

$$\mathcal{L} = \frac{-1}{B}\sum_{i=1}^{B}\sum_{j=1}^{l}\log P\left(y_j^{(i)} \mid y_{<j}^{(i)}, \mathcal{M}^a(x^{(i)})\right) \cdot I_{y_j^{(i)} \neq -100}, \tag{5}$$

where B denotes the batch size; l denotes the target sequence length of the i-th sample; P denotes the predicted probability of $y_j^{(i)}$ at decoding step j based on $y_{<j}^{(i)}$; $I_{y_j^{(i)} \neq -100}$ indicates that only tokens whose labels are not equal to -100 (i.e., not masked) participate in the loss.

The overall loss function for training the assistant model is formulated as follows:

$$\mathcal{L}_{multi}^a = \lambda_{cls}^a \cdot \mathcal{L}_{cls}^a + \lambda_{rea}^a \cdot \mathcal{L}_{rea}^a, \tag{6}$$

where λ_{cls}^a and λ_{rea}^a are the weighting hyperparameters to ensure a balanced trade-off between two tasks. After training, we can obtain the trained assistant model, $\overline{M}^a$.

Regarding data augmentation, given the limited capabilities of the assistant model, we only retain training samples for which sentiment can be correctly predicted through sentiment reasoning. See the material for more details.

$$\mathcal{D}_{rea}^a = \{(x_i, \widehat{c}_i^a, \widehat{y}_i^a) \mid \widehat{y}_i^a = L_i\}_{i=1}^{N_a}, \tag{7}$$

where $\widehat{c}_i^a, \widehat{y}_i^a = \overline{\mathcal{M}}^a\left(x_i; \mathcal{T}_{\text{pre}}\right)$ and $N_a < N$.

The complete sentiment reasoning dataset is obtained, which is used to train a student model.

$$\mathcal{D}_{rea}^{all} = \mathcal{D}_{rea}^{t} \cup \mathcal{D}_{rea}^{a}. \tag{8}$$

3.3.2 Student Model with Joint Learning

To enable efficient deployment in resource-constrained environments, we employ a lightweight student MLLM, $\mathcal{M}^s$, trained through knowledge distillation. The student model jointly learns from two sources, including ground-truth labels (hard labels) for accurate prediction and probability distributions (soft labels) from the assistant model to capture its reasoning patterns. The dual supervision allows the student model to inherit the assistant model's discriminative capabilities.

Hard Label. The student model undergoes fine-tuning using constructed reasoning data, $\mathcal{D}_{rea}^{all}$, enabling it to acquire step-by-step reasoning capabilities through reasoning distillation. The hard label loss is defined as follows:

$$\begin{cases} \mathcal{L}_{cls}^{s_{hard}} = \mathbb{E}_{\mathcal{D}_{rea}^{all}} \log P\left([x; L] \mid \mathcal{M}^s\right) \\ \mathcal{L}_{rea}^{s_{hard}} = \mathbb{E}_{\mathcal{D}_{rea}^{all}} \log P\left([x; c] \mid \mathcal{M}^s\right), \end{cases} \tag{9}$$

where P denotes the probability distribution; c represents the reasoning process. The losses $\mathcal{L}_{cls}^{s_{hard}}$ and $\mathcal{L}_{rea}^{s_{hard}}$ are used to train the student model to learn the direct mapping from multimodal input to sentiment labels and to generate coherent sentiment reasoning, respectively.

Soft Label. To address the black-box nature of closed-source MLLMs, the assistant model is employed as an intermediary to provide soft labels for distillation. Given an input x, the probability distribution p_k at the k-th position is obtained from the logit value z_k through a single forward pass followed by the softmax function. It is formally defined as:

$$p_k = \frac{\exp\left(z_k/\tau\right)}{\sum_j exp\left(z_j/\tau\right)}, \tag{10}$$

where τ denotes the temperature hyperparameter, which is used to control the smoothness of the distribution.

After obtaining the probability distributions p^a from $\mathcal{M}^a$ and p^s from $\mathcal{M}^s$, we employ the Kullback–Leibler (KL) (Wu et al., 2025) divergence to minimize the discrepancy between the two distributions. It enables the student model to mimic the prediction behavior of the larger model. The training for soft label distillation is defined as follows:

$$\begin{cases} \mathcal{L}_{soft}\left(p^a, p^s\right) = \sum_k p_k^a \log \frac{p_k^a}{p_k^s} \\ \mathcal{L}_{cls}^{s_{soft}} = \mathcal{L}_{soft}\left(p_{cls}^a, p_{cls}^s\right) \\ \mathcal{L}_{rea}^{s_{soft}} = \mathcal{L}_{soft}\left(p_{rea}^a, p_{rea}^s\right). \end{cases} \tag{11}$$

Joint Learning. The student model training retains the multi-task learning. The overall hard-label loss and soft-label loss for the student model are defined as follows:

$$\begin{cases} \mathcal{L}_{multi}^{s_{hard}} = \lambda_{cls}^{s_{hard}} \cdot \mathcal{L}_{cls}^{s_{hard}} + \lambda_{rea}^{s_{hard}} \cdot \mathcal{L}_{rea}^{s_{hard}} \\ \mathcal{L}_{multi}^{s_{soft}} = \lambda_{cls}^{s_{soft}} \cdot \mathcal{L}_{cls}^{s_{soft}} + \lambda_{rea}^{s_{soft}} \cdot \mathcal{L}_{rea}^{s_{soft}} \end{cases} \tag{12}$$

where $\lambda_{cls}^{s_{hard}}$, $\lambda_{rea}^{s_{hard}}$, $\lambda_{cls}^{s_{soft}}$, and $\lambda_{rea}^{s_{soft}}$ are hyperparameters that balance the contributions of classification loss and reasoning generation loss in the hard-label and soft-label multi-task learning objectives, respectively.

To jointly leverage hard-label and soft-label supervision, we define the total loss of the student model as follows.

$$\mathcal{L}_{total}^{s} = (1-\lambda)\,\mathcal{L}_{multi}^{s_{hard}} + \lambda \mathcal{L}_{multi}^{s_{soft}}, \tag{13}$$

where λ is a hyperparameter that controls the balance between hard-label and soft-label supervision.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on both coarse-grained MSA, MVSA-Single and MVSA-Multiple datasets, preprocessed following (Liu et al., 2024) and fine-grained MSA, Twitter-2015 and Twitter-2017 datasets (Yu and Jiang, 2019). Table 1 presents the statistics of four datasets with the constructed sentiment reasoning data for JMSRC.

Dataset	Train	Dev	Test	Train^{g+}	Train^{q+}
MVSA-Single	3608	451	452	6483	6350
MVSA-Multiple	13619	1702	1702	23424	23697
Twitter-2015	3179	1122	1037	6166	6218
Twitter-2017	3562	1176	1234	6652	6871

Table 1: Statistics of datasets. $g+$ and $q+$ represent the teacher models GPT-4o-mini (Hurst et al., 2024) and Qwen2.5-VL-72B (Bai et al., 2025), respectively.

4.1.2 Model Selection

To build an efficient hierarchical reasoning distillation, we design four distillation architectures, as summarized in Table 2. Note that, while our model selection is limited, experimental results clearly demonstrate the effectiveness of MulCoT-RD. See the Appendix B for more details.

ID	Teacher Model	Assistant Model	Student Model
1	GPT-4o-mini	Qwen3-VL-8B	Qwen3-VL-2B
2		Qwen2.5-VL-7B	Qwen2.5-VL-3B
3	Qwen2.5-VL-72B	Qwen3-VL-8B	Qwen3-VL-2B
4		Qwen2.5-VL-7B	Qwen2.5-VL-3B

Table 2: Four reasoning distillation architectures.

4.1.3 Implementation Details

We train our models on NVIDIA RTX A6000 GPUs using the AdamW optimizer (Loshchilov and Hutter, 2017). During training, we set the initial learning rate to 3e-4 and employ a dynamic adjustment strategy: if the validation set performance does not improve for two consecutive epochs, we halve the learning rate until it reaches a minimum of 1e-6. Due to resource limitations, we set the batch size to 2 and train for a maximum of 20 epochs. To mitigate instability caused by small batch sizes, we use gradient accumulation, updating parameters every 20 steps. The multi-task learning hyperparameters λ_{rea}^{a}, $\lambda_{rea}^{s_{hard}}$, $\lambda_{rea}^{s_{soft}}$ and λ_{cls}^{a}, $\lambda_{cls}^{s_{hard}}$, $\lambda_{cls}^{s_{soft}}$ are set to 0.8 and 0.2, respectively, while the knowledge distillation coefficient λ is set to 0.3. Detailed explanations and configurations can be found in Appendix D

4.1.4 Evaluation Metrics

In line with previous work (Chen et al., 2024), we evaluate model performance of classification on coarse-grained MSA using Accuracy (**Acc**) and Weighted F1 (**w-F1**). For fine-grained MSA (MASC), we follow previous studies (Zhou et al., 2023) and adopt Accuracy and Macro F1 (**m-F1**) as evaluation metrics. For the sentiment reasoning task, we employ comprehensive metrics including sentence embedding-based cosine similarity (**Sim**) (Reimers and Gurevych, 2019), **METEOR** (Banerjee and Lavie, 2005), **BLEU** (Papineni et al., 2002), **ROUGE-L** (Lin, 2004), and Distinct-N1/N2 (**Dist-1/2**) (Li et al., 2015).

4.2 Baselines

We compare popular models on **coarse-grained MSA** with MulCoT-RD, including **MultiSentiNet** (Xu and Mao, 2017), **HSAN** (Xu, 2017), **CoMN-Hop6** (Xu et al., 2018), **MGNNS** (Yang et al., 2021), **CLMLF** (Li et al., 2022), **MVCN** (Wei et al., 2023), $\mathbf{D^2R}$ (Chen et al., 2024). For **fine-grained MSA**, involving **ESAFN** (Yu et al., 2019), **TomBERT** (Yu and Jiang, 2019), **CapTrBERT** (Khan and Fu, 2021), **JML** (Ju et al., 2021), **VLP-MABSA** (Ling et al., 2022), **CMMT** (Yang et al., 2022), **AoM** (Zhou et al., 2023), **AETS** (Zhu et al., 2025). **Emotion-LLaMA** (Cheng et al., 2024) employs pretraining and instruction tuning based on LLaMA2-7B-Chat to enhance multimodal emotion recognition and explanation. **Qwen3-VL-8B-Thinking** (Yang et al., 2025a) features Interleaved-MRoPE and DeepStack for powerful spatial-temporal reasoning. Detailed descriptions can be found in Appendix C.

4.3 Main Results

Unlike previous models that only perform multimodal sentiment classification, our model enables joint sentiment reasoning and classification. We conduct experiments on both multimodal sentiment classification and reasoning tasks.

4.3.1 Results of Multimodal Sentiment Classification

Performance on coarse-grained MSA. Table 3 presents the comparison results on the coarse-grained MSA task. MulCoT-RD outperforms both the second-best model (Emotion-LLaMA) and the previous state-of-the-art model ($\mathbf{D^2R}$) on the MVSA-Single and MVSA-Multiple datasets, achieving substantial improvements. It highlights the benefits of explicitly modeling intra-modal sentiment structures and cross-modal reasoning processes. Notably, although the teacher model has greater parameter capacity, its lack of task-specific fine-tuning for MSA leads to suboptimal modeling of cross-modal emotional relations, making it inferior to the assistant model optimized with task-oriented objectives. Moreover, the student model outperforms the assistant model in certain cases, likely due to benefiting from the augmented training data generated by the assistant, which improves its generalization and robustness.

Performance on MASC. As shown in Table 4, the MulCoT-RD(asst) model (with Qwen2.5-VL-72B as the teacher) achieves the best overall performance. Compared to the second-best models AoM and AETS, MulCoT-RD(asst) exhibits a slight decrease in accuracy on the Twitter-2017 dataset

Model	Venue	MVSA-S		MVSA-M	
		Acc	w-F1	Acc	w-F1
MultiSentiNet	CIKM'17	69.8	69.8	68.9	68.1
HSAN	ISI'17	69.9	66.9	68.0	67.8
CoMN-Hop6	SIGIR'18	70.5	70.0	68.9	68.8
MGNNS	ACL'21	73.8	72.7	72.5	69.3
CLMLF	NAACL'21	75.3	73.5	72.0	69.8
MVCN	ACL'23	76.1	74.6	72.1	70.0
$\mathbf{D^2R}$	EMNLP'24	76.7	75.6	71.6	70.9
Emotion-LLaMA†	NeurIPS'24	82.7	81.8	75.6	**75.2**
Open-Flamingo‡	ICML'25	66.3	-	68.7	-
Qwen2.5-VL-3B*	Student	62.8	66.4	74.2	70.7
Qwen2.5-VL-7B*	Assistant	67.7	69.6	74.7	70.9
3-VL-8B-Thinking*		72.3	71.5	70.2	69.4
GPT-4o-mini*	Teacher[1]	76.7	75.6	71.6	71.4
MulCoT-RD(asst)		**83.6**	82.8	75.7	72.9
MulCoT-RD(stu)		82.7	82.3	76.9	74.2
Qwen2.5-VL-72B*	Teacher[2]	67.9	70.8	74.2	71.8
MulCoT-RD(asst)		83.2	82.1	76.9	73.8
MulCoT-RD(stu)		83.4	**83.2**	**77.2**	74.4

Table 3: Results for coarse-grained MSA. Models above the middle line are small models fully fine-tuned, while those below are (M)LLMs fine-tuned with LoRA. † denotes the results reproduced by us using models retrained on our datasets. ‡ indicates the 16-shot performance under In-Context Learning (ICL). The best results are bold-typed and the second best ones are underlined. * means the zero-shot performance.

Model	Venue	Twitter-15		Twitter-17	
		Acc	m-F1	Acc	m-F1
ESAFN	TASLP'20	73.4	67.4	67.8	64.2
TomBERT	IJCAI'19	77.2	71.8	70.5	68.0
CapTrBERT	ACM MM'21	78.0	73.2	72.3	70.2
JML	EMNLP'21	78.7	-	72.7	-
VLP-MABSA	ACL'22	78.6	73.8	73.8	71.8
CMMT	IPM'22	77.9	-	73.8	-
AoM	ACL'23	80.2	75.9	76.4	75.0
AETS	AAAI'25	79.5	-	**76.6**	-
Emotion-LLaMA†	NeurIPS'24	73.9	70.2	69.2	67.9
Open-Flamingo‡	ICML'25	70.4	-	62.6	-
Qwen2.5-VL-3B*	Student	48.9	49.7	56.8	55.6
Qwen2.5-VL-7B*	Assistant	58.3	55.6	58.6	57.6
3-VL-8B-Thinking*		59.2	55.8	60.4	58.9
GPT-4o-mini*	Teacher[1]	49.4	37.6	54.0	52.8
MulCoT-RD(asst)		80.7	75.3	74.6	74.6
MulCoT-RD(stu)		80.4	75.2	74.0	73.3
Qwen2.5-VL-72B*	Teacher[2]	59.5	57.1	63.9	63.4
MulCoT-RD(asst)		**80.8**	**77.2**	75.0	**75.1**
MulCoT-RD(stu)		80.5	75.1	74.3	74.1

Table 4: Results of different methods for MASC. "-" means it does not exist in the original paper.

by 1.4% and 1.6%, respectively, but consistently achieves the highest scores across all other evaluation metrics. We attribute this to two primary reasons. First, the Twitter-2017 dataset contains a large number of unparseable and unrecognizable symbols (Peng et al., 2024), including emojis that are commonly used on Twitter. These symbols may mislead the model by obscuring emotional semantics during reasoning, thereby slightly reducing accuracy. Second, MulCoT-RD(asst) is fine-tuned using LoRA, whereas most existing SOTA methods, such as AoM and AETS, adopt full-parameter fine-tuning. This limits the extent of parameter updates during task adaptation, resulting in smaller performance gains compared to full fine-tuning (Biderman et al., 2024). Given this, we believe our proposed method remains effective for MASC.

Notably, the student model of MulCoT-RD contains only 3B parameters, significantly fewer than the large multimodal architecture of Emotion-LLaMA (Cheng et al., 2024), which combines LLaMA2-7B-chat with encoders like EVA, CLIP, VideoMAE, and HuBERT-large. Despite its smaller size, MulCoT-RD(stu) outperforms Emotion-LLaMA on multiple benchmarks, demonstrating superior efficiency and strong applicability in resource-constrained settings.

4.3.2 Evaluation of Sentiment Reasoning

MulCoT-RD achieves efficient and effective sentiment reasoning. We evaluate the reasoning performance of the student and assistant models, as well as Emotion-LLaMA, using the sentiment reasoning process from the teacher model as gold-standard references (exemplified by GPT-4o-mini), with results presented in Table 5. Our models achieve a comprehensive performance advantage over Emotion-LLaMA across all key reasoning metrics. The results demonstrate high-quality sentiment reasoning generation across multiple evaluation metrics. Cosine similarity (Sim) consistently exceeds 90% across all models, confirming strong semantic alignment between generated and gold-standard reasoning chains. METEOR scores ranging from 45.4% to 59.8% further indicate substantial paraphrase-level and lexical overlap. While BLEU and ROUGE-L show some fluctuations, coarse-grained MSA variants generally outperform fine-grained MSA, reflecting better surface-form alignment. Distinct-N1 and Distinct-N2 scores remain approximately 49% and 80%, respectively, indicating that the generated reasoning maintains high linguistic diversity, enhancing the interpretability and robustness of reasoning tasks.

4.4 Ablation Study

In this section, we investigate the impact of each MulCoT-RD component, with results presented in Table 6. When we only use the text modality (**w/o Img**), the model performs worse on all metrics

Model	Dataset	Sim	Meteor	Bleu	Rouge-L	Dist-1	Dist-2
ELLA	MVSA-S	87.6	35.9	14.6	35.1	49.8	80.2
	MVSA-M	84.7	36.0	15.9	35.9	52.5	83.7
	Twitter-15	86.3	38.6	18.3	39.3	42.7	72.9
	Twitter-17	86.6	38.1	17.6	38.2	43.0	73.1
Asst	MVSA-S	92.6	59.8	47.8	55.0	49.8	80.2
	MVSA-M	93.0	57.4	48.1	57.2	48.6	79.4
	Twitter-15	92.9	54.6	43.0	58.3	42.4	72.9
	Twitter-17	90.5	51.2	35.9	53.3	45.2	74.1
Stu	MVSA-S	92.2	47.3	58.8	54.2	49.8	80.2
	MVSA-M	92.1	56.8	46.7	55.8	49.5	80.3
	Twitter-15	90.3	45.4	28.2	46.0	49.5	79.9
	Twitter-17	90.0	49.2	33.1	50.8	45.2	74.1

Table 5: Evaluation results of generated reasoning from Emotion-LLaMA, assistant and student models.

compared to the complete model, highlighting the importance of incorporating visual modality. Similarly, when we remove the text modality (**w/o Text**), the model has a significant performance drop on all datasets. The decline, more severe than w/o Img, highlights the key role of text and the necessity of multimodal integration. **w/o CoT** means to remove the multi-task learning paradigm and exclude the sentiment reasoning task from the training process, leading to a general performance drop. It highlights the importance of deeply modeling intra-modal and cross-modal sentiment reasoning. Note that all the above ablation experiments are conducted on the assistant model. **w/o Asst** omits the assistant model, removing the use of soft labels in the distillation process and reducing the scale and diversity of training data. This leads to a notable performance drop across all datasets, demonstrating the effectiveness of the teacher–assistant–student hierarchical distillation framework for JMSRC.

Method	MVSA-S Acc	MVSA-S w-F1	MVSA-M Acc	MVSA-M w-F1	Twitter-15 Acc	Twitter-15 m-F1	Twitter-17 Acc	Twitter-17 w-F1
MulCoT-RD	**83.2**	**82.1**	**76.9**	73.8	**80.8**	**77.2**	**75.0**	**75.1**
w/o Img	79.4	77.7	73.7	73.0	78.4	72.5	73.5	73.5
w/o Txt	77.9	77.1	66.2	67.7	65.6	56.6	64.6	59.4
w/o CoT	79.9	79.7	74.2	73.1	79.9	75.5	74.2	73.4
w/o Asst	81.9	81.3	75.2	**74.1**	79.3	72.3	73.7	73.3

Table 6: The performance comparison of our full model and its ablated methods under the setting where Qwen2.5-VL-72B serves as the teacher model.

4.5 Efficiency of MulCoT-RD

To further demonstrate the practicability of MulCoT-RD, we provide the model efficiency comparison in Figure 3. We find that, on the MVSA-Single and Twitter-2015 datasets, our distilled student models (Qwen2.5-VL-3B and Qwen3-VL-2B) achieve significantly lower inference latency and GPU memory usage than Emotion-LLaMA, while obtaining notable improvements in Accuracy. Specifically, in Twitter-2015, Qwen3-VL-2B achieves the 8.79% accuracy improvement (80.4 vs. 73.9) with 5.81x fewer parameters and 0.33x faster inference speed compared to Emotion-LLaMA.

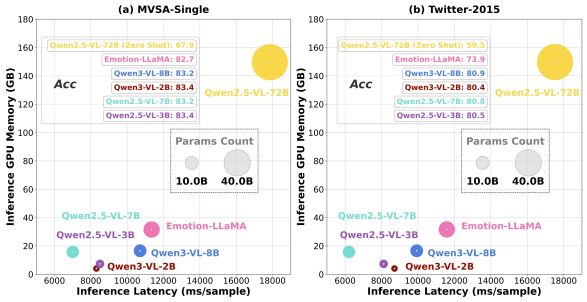

Figure 3: Efficiency comparison on MVSA-Single and Twitter-2015. Metrics are measured on the batch size as 1 and all samples are from the test set. Note that the models are trained under the paradigm where Qwen2.5-VL-72B serves as the teacher model. Detailed efficiency comparison results are provided in Appendix H.

4.6 Robustness of MulCoT-RD

To validate the robustness of our approach across different backbones, we conducted the base-model adaptation study by replacing the Qwen2.5-VL series with the Qwen3-VL series (as shown in Table 7) and the Flan-T5 series (see Appendix G). The results demonstrate that the models maintain strong performance across different backbones, illustrating the robustness and adaptability of MulCoT-RD.

Qwen3-VL	MVSA-S Acc	MVSA-S w-F1	MVSA-M Acc	MVSA-M w-F1	Twitter-15 Acc	Twitter-15 m-F1	Twitter-17 Acc	Twitter-17 m-F1
8B (asst)[1]	83.6	82.9	76.0	73.5	80.5	76.3	**76.1**	**75.1**
2B (stu)[1]	**84.1**	**83.7**	76.7	74.1	80.0	75.2	74.8	74.7
8B (asst)[2]	83.2	82.6	76.7	73.2	**80.9**	**77.4**	75.0	73.4
2B (stu)[2]	83.4	82.4	**76.8**	**74.2**	80.4	76.7	73.8	73.0

Table 7: Performance of Qwen3-VL-based models on coarse-grained MSA and MASC. The best results are bold-typed, and the second best ones are underlined. 1 and 2 indicate the Teacher models, with 1 being GPT-4o-mini and 2 being Qwen2.5-VL-72B.

4.7 Case Study

To validate MulCoT-RD's effectiveness, we present two illustrative cases in Figure 4. In case (a), $\mathbf{D^2R}$ incorrectly predicts sentiment by overrelying on surface-level positive terms like "popular" and "bipartisan" while missing the emotional shift from the word "hopeless" which establishes a negative tone. MulCoT-RD successfully captures this reversal. In case (b), the AoM misclassifies sentiment

Task	Coarse-Grained MSA	Fine-Grained MSA
Image		
Text	**(a)** RT @nationaljournal: Supreme Court term limits are popular, bipartisan, and hopeless.	**(b)** RT @ tunkuv : Wedding photo becomes symbol of South Korea **MERS** scare
Previous SOTA	**Method:** D^2R **Prediction:** Positive ✗	**Method:** AoM **Prediction:** Negative ✗
MulCoT-RD (ours)	***Sentiment Reasoning*** **"Text_analysis":** "The text discusses the popularity and bipartisan support of Supreme Court term limits but calls them 'hopeless' indicating negative sentiment about feasibility." **"Image_analysis":** "The image shows a large crowd outside the Supreme Court, suggesting public concern and urgency." **"Conflict_resolution":** "Despite some positive points, the overall tone is negative due to 'hopeless.' The crowd image reinforces public frustration or disappointment." **"Final_conclusion":** "The text's negative view aligns with the image's portrayal of public concern, jointly conveying an overall negative sentiment."	***Sentiment Reasoning*** **"Text_analysis":** "The text says the wedding photo symbolizes the South Korea MERS scare but does not express positive or negative sentiment toward MERS, focusing on cultural significance with a neutral tone." **"Image_analysis":** "The image shows people, including a newlywed couple, wearing masks possibly related to MERS, but it conveys no clear sentiment and maintains a neutral tone." **"Conflict_resolution":** "There is no conflict; both text and image present neutral, factual observations without emotional expression." **"Final_conclusion":** "Overall, the text and image provide objective, neutral information about the symbolic meaning and health measures, supporting a neutral classification."
	"Prediction": "Negative" ***Sentiment Classification*** ✓	**"Prediction":** "Neutral" ***Sentiment Classification*** ✓

Figure 4: Visualization of two samples, using the MulCoT-RD architecture with ID 4 from Table 2.

for the aspect term "MERS" by focusing on superficially negative words like "scare", leading to misinterpretation. MulCoT-RD effectively distinguishes between author stance (factual reporting) and content sentiment, producing correct predictions. This superior performance stems from our multi-task learning mechanism that integrates CoT reasoning and sentiment classification, enabling comprehensive modeling of intra-modal and cross-modal sentiment reasoning.

5 Conclusion

We focus on Joint Multimodal Sentiment Reasoning and Classification, JMSRC, in the resource-limited scenario that simultaneously generates multimodal reasoning chains and sentiment predictions. To address the dual challenges of reasoning interpretability and efficient deployment, we introduce MulCoT-RD, a unified framework combining structured CoT enhancement with reasoning distillation. Through a hierarchical teacher-assistant-student paradigm and joint multi-task learning, our method enables lightweight models to autonomously perform high-quality sentiment reasoning and classification. Extensive experiments across four datasets demonstrate the effectiveness and robustness of MulCoT-RD. In future work, we plan to incorporate direct preference optimization (DPO) with high- and low-quality reasoning sample filtering to further enhance the model's emotional reasoning quality and classification performance.

Limitations

MulCoT-RD demonstrates strong performance on joint multimodal sentiment reasoning and classification with a lightweight model, yet several limitations exist. Our approach depends on high-performance teacher models (e.g., GPT-4o-mini) to synthesize the initial reasoning dataset via carefully crafted Chain-of-Thought prompts. This generation process entails considerable inference costs and is sensitive to the consistency and quality of

the teacher model's outputs. The evaluation of generated reasoning chains relies mainly on automatic metrics, including n-gram overlap measures and embedding-based cosine similarity. These metrics may fail to adequately assess semantic coherence, factual accuracy, logical consistency, or human-perceived interpretability. Due to resource constraints, no large-scale human evaluation of reasoning quality was conducted. Although the student model is lightweight and targeted at resource-constrained deployment, it is built on a vision-language architecture that requires multimodal processing capabilities. This results in higher memory and computational demands than unimodal text-only models, which may limit deployment on extremely low-resource devices with minimal hardware acceleration support. We leave the exploration of more cost-effective reasoning data generation, comprehensive human evaluation of reasoning quality, multilingual generalization, and deployment on edge devices for future work.

Ethical considerations

This work constructs a multimodal sentiment reasoning dataset by augmenting existing publicly available multimodal sentiment analysis datasets with model-generated reasoning produced by closed-source large multimodal language models (MLLMs). All source datasets are publicly available and are used in accordance with their original licenses and terms of use. The closed-source MLLMs are accessed exclusively through official APIs, and no private, proprietary, or user-identifiable data are accessed or collected.

The proposed dataset does not introduce new personal information beyond what is already present in the original public datasets, nor does it attempt to infer sensitive personal attributes. We acknowledge that sentiment interpretation is inherently subjective and that model-generated reasoning may reflect biases present in the underlying models. To reduce this risk, we employ structured prompting and verification procedures to improve annotation consistency and reliability.

The dataset is intended solely for research purposes in multimodal sentiment understanding and reasoning. We believe that this work adheres to established ethical standards for data usage and responsible application of large-scale models, and we do not anticipate foreseeable misuse beyond the scope of existing sentiment analysis research.

Acknowledgements

The work is supported by the Fundamental Research Funds for the Central Universities under Grants (N25XQD004, N25ZLL045), and the National Natural Science Foundation of China (Nos. 62272092, 62172086, 62576085).

References

Shahin Amiriparian, Lukas Christ, Alexander Kathan, Maurice Gerczuk, Niklas Müller, Steffen Klug, Lukas Stappen, Andreas König, Erik Cambria, Björn W Schuller, and 1 others. 2024. The muse 2024 multimodal sentiment analysis challenge: Social perception and humor recognition. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*, pages 1–9.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, and 1 others. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. 2024. D2r: Dual-branch dynamic routing network for multimodal sentiment detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3536–3547.

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

Li Chenglin, Qianglong Chen, Liangyue Li, Caiyu Wang, Feng Tao, Yicheng Li, Zulong Chen, and Yin Zhang. 2024. Mixed distillation helps smaller language models reason better. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1673–1690.

Yanqi Dai, Zebin You, Dong Jing, Yutian Luo, Nanyi Fei, Guoxing Yang, and Zhiwu Lu. 2024. Cotbal: Comprehensive task balancing for multi-task visual instruction tuning. *arXiv preprint arXiv:2403.04343*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. 2020. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE.

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*.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4395–4405.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042.

Junho Kim, Jun-Hyung Park, Mingyu Lee, Wing-Lam Mok, Joon-Young Choi, and SangKeun Lee. 2022. Tutoring helps students learn better: Improving knowledge distillation for bert with tutor network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7382.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-kd: Making small language models better multi-step reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17643–17658.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*.

Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. 2025a. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–23.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025b. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07955*.

Wuchao Liu, Wengen Li, Yu-Ping Ruan, Yulou Shu, Juntao Chen, Yina Li, Caili Yu, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. 2024. Weakly correlated multimodal sentiment analysis: New dataset and topic-oriented model. *IEEE Transactions on Affective Computing*, 15(4):2070–2082.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. 2023. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tianshuo Peng, Zuchao Li, Ping Wang, Lefei Zhang, and Hai Zhao. 2024. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In *Proceedings of the AAAI Conference*

on Artificial Intelligence, volume 38, pages 18869–18878.

Pol G Recasens, Ferran Agullo, Yue Zhu, Chen Wang, Eun Kyung Lee, Olivier Tardieu, Jordi Torres, and Josep Ll Berral. 2025. Mind the memory gap: Unveiling gpu bottlenecks in large-batch llm inference. *arXiv preprint arXiv:2503.08311*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

OpenBMB MiniCPM-o Team. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone.

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

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020a. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

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.

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

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. 2025. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508.

Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE international conference on intelligence and security informatics (ISI)*, pages 152–154. IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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*.

Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*, 59(5):103038.

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

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. 2025b. Learning to rebalance multimodal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025. Harnessing multimodal large language models for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13069–13077.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767.

Yazhou Zhang, Prayag Tiwari, Lu Rong, Rui Chen, Nojoom A AlNajem, and M Shamim Hossain. 2022. Affective interaction: Attentive representation learning for multi-modal sentiment classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 18(3s):1–23.

Yiqun Zhang, Xiaocui Yang, Xiaobai Li, Siyuan Yu, Yi Luan, Shi Feng, Daling Wang, and Yifei Zhang. 2024. Psydraw: A multi-agent multimodal system for mental health screening in left-behind children. *arXiv preprint arXiv:2412.14769*.

Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8184–8196.

Linlin Zhu, Heli Sun, Qunshu Gao, Yuze Liu, and Liang He. 2025. Aspect enhancement and text simplification in multimodal aspect-based sentiment analysis for multi-aspect and multi-sentiment scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1683–1691.

A Data expansion with Assistant Model

After training the assistant model, we apply it to perform inference on the **original training set only**, explicitly excluding the validation and test sets to prevent any risk of label leakage. During this process, we retain only those samples whose predicted sentiment labels match the ground truth. These correctly predicted samples are then merged with the original training set to construct an expanded dataset, which is subsequently used for training the student model. Detailed results of the data expansion are presented in Table 8.

This strategy significantly increases the scale and diversity of the training data, broadens the coverage of sentiment label distributions, and incurs no additional manual annotation cost. It equips the student model with richer and higher-quality learning signals, effectively mitigating the challenge of limited annotated data commonly encountered in multimodal sentiment analysis tasks.

Dataset	Samples	GPT-4o-mini			Qwen2.5-VL-72B		
		Acc	w-F1	m-F1	Acc	w-F1	m-F1
MVSA-S	3608	79.7	79.7	69.9	76.0	77.1	66.9
MVSA-M	13619	72.0	68.0	55.3	74.0	70.5	60.6
Twitter-15	3179	94.0	94.1	92.6	95.6	95.6	94.6
Twitter-17	3562	86.8	86.7	86.4	92.9	92.9	93.4

Table 8: Performance of the Assistant Model (Qwen-2.5-VL-7B) on Training Sets During Data Expansion, Guided by Different Teacher Models.

B Model Selection

To construct a hierarchical reasoning distillation framework for achieving efficient joint multimodal sentiment reasoning and classification (JMSRC), we carefully select the following models as the teacher model, the assistant model, and the student model. Table 9 shows the specific model selections and their characteristics.

Role	Model	Access	Release Date
Teacher	GPT-4o-mini	Closed	2024.07
	Qwen2.5-VL-72B	Open	2025.02
Assistant	Qwen3-VL-8B	Open	2025.10
	Qwen2.5-VL-7B	Open	2025.02
Student	Qwen3-VL-2B	Open	2025.10
	Qwen2.5-VL-3B	Open	2025.02

Table 9: Model Selection and Characteristics.

C Baselines

Methods for coarse-grained MSA. 1) **MultiSentiNet** (Xu and Mao, 2017) is a deep attention-based semantic network for multimodal sentiment analysis. 2) **HSAN** (Xu, 2017) is a hierarchical semantic attentional network based on image captions for multimodal sentiment analysis. 3) **CoMN-Hop6** (Xu et al., 2018) utilizes co-memory network to iteratively model the interactions between multiple modalities. 4) **MGNNS** (Yang et al., 2021) adopts multi-channel graph neural networks with sentiment-awareness for image-text sentiment detection. 5) **CLMLF** (Li et al., 2022) proposes a contrastive learning and multi-layer fusion method for multimodal sentiment detection. 6) **MVCN** (Wei et al., 2023) designs a multi-view calibration network to solve the modality heterogeneity for multimodal sentiment detection. 7) **D^{2}R** (Chen

et al., 2024) proposes a dual-branch dynamic routing network to enhance multimodal sentiment detection by effectively modeling cross-modal interactions. 8) **Emotion-LLaMA** (Cheng et al., 2024) employs a specialized emotion tokenizer and instruction fine-tuning based on the LLaMA2-7B-chat to enhance multimodal emotion recognition. 9) **Qwen3-VL-8B-Thinking** (Yang et al., 2025a) features Interleaved-MRoPE and DeepStack for powerful spatial-temporal reasoning, plus precise Text–Timestamp Alignment. With native 256K context, it excels at complex STEM and logic tasks.

Methods for fine-grained MSA. 1) **ESAFN** (Yu et al., 2019) is an entity-level sentiment analysis method based on LSTM. 2) **TomBERT** (Yu and Jiang, 2019) applies BERT to obtain aspect-sensitive textual representations. 3) **CapTrBERT** (Khan and Fu, 2021) translates images into text and construct an auxiliary sentence for fusion. 4) **JML** (Ju et al., 2021) is the first joint model for MABSA with an auxiliary cross-modal relation detection module. 5) **VLP-MABSA** (Ling et al., 2022) performs five task-specific pretraining tasks to model aspects, opinions, and alignments. 6) **CMMT** (Yang et al., 2022) implements a gate to control the multimodal information contributions during inter-modal interactions. 7) **AoM** (Zhou et al., 2023) introduces an aspect-oriented network designed to reduce visual and textual distractions from complex image-text interactions. 8) **Emotion-LLaMA** (Cheng et al., 2024). 9) **AETS** (Zhu et al., 2025) improves multimodal sentiment analysis by enhancing aspects and simplifying text. 10) **Qwen3-VL-8B-Thinking** (Yang et al., 2025a)

D Implementation Details

D.1 Hyperparameters in Multi-Task Learning

In our multi-task learning setup, we assign weights of 0.8 and 0.2 to the CoT (Chain-of-Thought) generation task and the sentiment classification task, respectively. This design is motivated by the following considerations:

- **Task complexity**: CoT generation involves structured reasoning and belongs to a class of complex sequence generation tasks, which are more difficult to train and typically incur higher loss values. In contrast, sentiment classification is a relatively simple three-way classification task. Therefore, assigning a higher weight to CoT generation encourages the model to focus more on learning reasoning capabilities.

- **Convergence and gradient sensitivity**: As shown in Figure 5, when the loss weights of the CoT generation task are set to 0.8 and 0.2, the model exhibits a significant difference in converged loss: specifically, the loss differs by approximately 4x on the MVSA-Single dataset (0.0134 vs. 0.0033) and by approximately 12x on the Twitter-2015 dataset (0.0084 vs. 0.0007). In contrast, for the sentiment classification task, the model's converged loss remains largely consistent under different loss weight settings. Preliminary experiments show that the CoT task converges more slowly and is more sensitive to gradient fluctuations. Increasing its loss weight helps amplify gradient signals and improves training stability and task performance.

- **Empirical validation**: We experimented with different weight configurations (e.g., {0.5, 0.5}, {0.2, 0.8}) and observed that assigning lower weights to the CoT task led to slower loss reduction and decreased classification accuracy. In contrast, the {0.8, 0.2} setting consistently yielded better performance on both the validation and test sets.

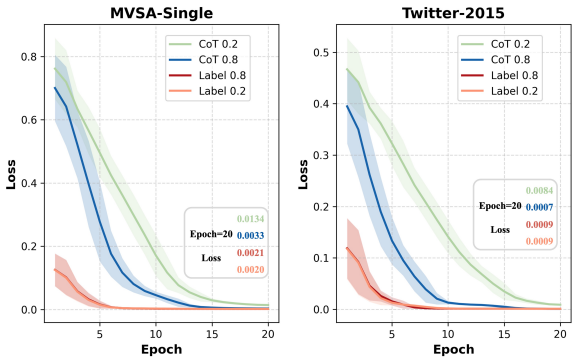

Figure 5: Effect of Loss Weight on Convergence for CoT Generation and Sentiment Classification Tasks.

This weighting scheme also reflects the task balancing principle proposed by CoTBal (Dai et al., 2024), which emphasizes that in multi-task scenarios, loss weights should be adaptively assigned based on task complexity and learning dynamics to enhance main-task optimization and overall model performance.

D.2 Hyperparameter in Knowledge Distillation

We set the hyperparameter λ to 0.3, following the empirical practices in prior work (Lee et al., 2024), which achieve a good balance between stable training and effective knowledge transfer from the teacher model.

D.3 LoRA Configuration

In all experiments, we adopt Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning of the multimodal models, rather than updating all parameters. In our implementation, the LoRA rank is set to $r = 16$ and the scaling factor is set to $\alpha = 32$. This strategy allows the models to adapt effectively to downstream tasks under limited computational resources, while significantly reducing the number of trainable parameters. The proportion of trainable parameters for each model is summarized in Table 10, and further implementation details are available in the GitHub repository.

Model	Total Parameters	Trainable	Ratio (%)
Qwen3-VL-8B	8,810,770,672	43,646,976	0.4954
Qwen3-VL-2B	2,144,964,608	17,432,576	0.8127
Qwen2.5-VL-7B	8,339,756,032	47,589,376	0.5706
Qwen2.5-VL-3B	3,791,775,744	37,152,768	0.9798

Table 10: Total and trainable parameter counts under LoRA fine-tuning.

E Details of the Adaptive Retry Controller (ARC)

Since large multimodal models are still prone to hallucinations in open-ended generation—such as producing content that deviates from the task specification or violating the required output format—the predicted sentiment label cannot always be reliably extracted from a single response. To mitigate this issue, we introduce an Adaptive Retry Controller (ARC) in the inference stage. Whenever the initially generated response does not conform to the expected format or fails to yield a valid sentiment label, ARC automatically triggers a retry process, prompting the model to regenerate the response for the same input sample. This process is repeated until a valid output is obtained or a predefined maximum number of retries is reached.

In our implementation, the maximum number of retries is set to three. This choice strikes a balance between improving label extraction robustness and controlling computational overhead. Empirically, approximately 79% of the samples obtain a valid prediction within one or two generations, while only a small fraction require the full retry budget. As a result, ARC substantially reduces cases of missing or incorrectly parsed predictions, leading to a more stable and reliable inference pipeline without incurring prohibitive additional cost.

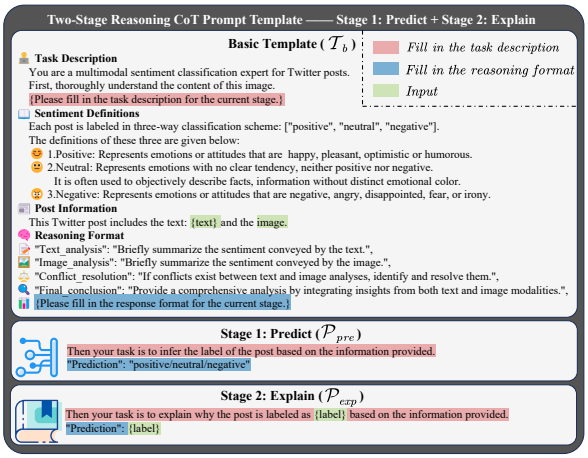

Figure 6: Two-stage reasoning prompt template.

F Two-stage reasoning prompt template

The prompt template for the two-stage multimodal Chain-of-Thought (CoT) is illustrated in Figure 6. By leveraging the basic template $\mathcal{T}_b$, the framework assigns a multimodal expert role to the model and prescribes a four-step structured reasoning process—comprising text analysis, image analysis, conflict resolution, and final conclusion—to explicitly address emotional correlations and conflicts between text and image modalities. Specifically, the execution is bifurcated into two stages: Prediction ($\mathcal{P}_{pre}$) and Explanation ($\mathcal{P}_{exp}$). In the first stage, the model is required to autonomously derive sentiment labels based on multimodal information to evaluate its inherent reasoning capabilities. In the second stage, the model is guided to backwardly construct logical justifications grounded in known labels, thereby generating deep-seated sentiment explanations. This two-stage decoupled design not only ensures the logical rigor of the reasoning chain but also substantially enhances the quality and interpretability of the generated data through the CoT mechanism.

G Robustness of MulCoT-RD

For the Flan-T5 series. We utilize MiniCPM-o-2.6 (Team, 2025) to generate image captions, con-

verting multimodal inputs to text-only format. Using the Flan-T5 architecture, we fine-tune both assistant and student models with full parameters, replicating the complete training pipeline including multimodal CoT enhancement, multi-task learning, and reasoning distillation. As shown in Figure 7 to 9, the Flan-T5-based models achieve strong performance despite having only 248M parameters, demonstrating the robustness and adaptability of MulCoT-RD across diverse backbone architectures.

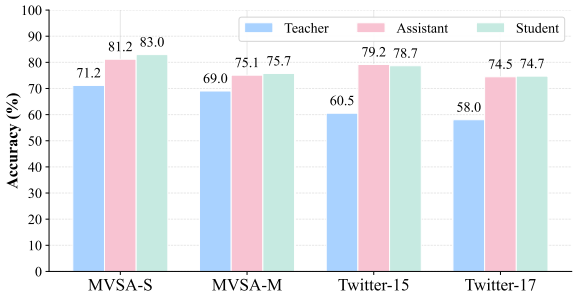

Figure 7: Accuracy comparison of teacher (GPT-3.5-Turbo), assistant (Flan-T5-Large with 783M parameters) and student (Flan-T5-Base) models.

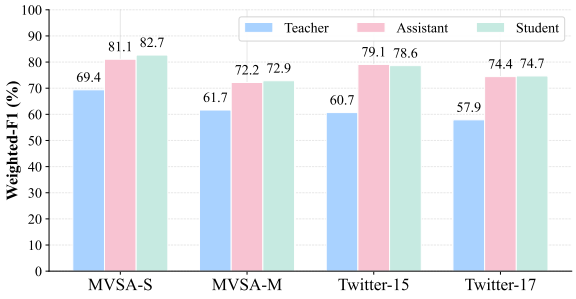

Figure 8: Weighted-F1 comparison of teacher(GPT-3.5-Turbo), assistant(Flan-T5-Large with 783M parameters) and student(Flan-T5-Base with 248M parameters) models.

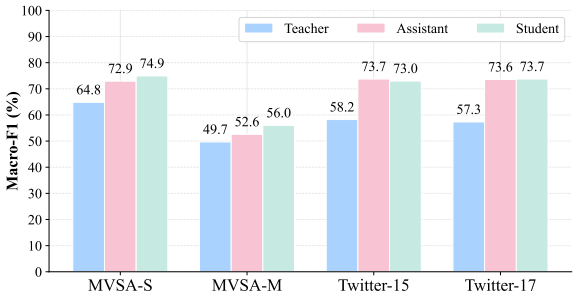

Figure 9: Macro-F1 comparison of teacher(GPT-3.5-Turbo), assistant(Flan-T5-Large with 783M parameters) and student(Flan-T5-Base with 248M parameters) models.

H Efficiency of MulCoT-RD

From the comparison of the data in Tables 11 and Table 12, Qwen3-VL-2B demonstrates strong competitiveness under resource-constrained scenarios. On the MVSA-Single dataset, its GPU memory consumption is only 4.11 GB, approximately 13% of Emotion-LLaMA (31.61 GB), substantially lowering the hardware deployment barrier. In contrast, our student models, by optimizing TFLOPs, achieve significantly reduced end-to-end inference latency while maintaining minimal memory usage. These results clearly indicate that knowledge distillation using high-quality two-stage CoT data enables lightweight models to achieve a favorable balance of high accuracy, low latency, and compact footprint in multimodal sentiment analysis tasks, providing reliable support for real-time online monitoring and mobile deployment.

Model	Params (B)	TFLOPs -	Latency (ms)	Memory (GB)
Qwen2.5-VL-72B	73.41	38.24	17849.37	149.69
Emotion-LLaMA	14.58	4.11	11342.73	31.61
Qwen3-VL-8B	8.81	3.13	10708.24	16.61
Qwen3-VL-2B	2.14	3.15	8302.95	4.11
Qwen2.5-VL-7B	8.34	3.02	7002.99	15.81
Qwen2.5-VL-3B	3.79	1.32	8500.58	7.35

Table 11: Efficiency comparison on the MVSA-Single dataset.

Model	Params (B)	TFLOPs -	Latency (ms)	Memory (GB)
Qwen2.5-VL-72B	73.41	38.41	17524.94	149.76
Emotion-LLaMA	14.58	4.31	11582.16	31.66
Qwen3-VL-8B	8.81	3.39	9930.38	16.62
Qwen3-VL-2B	2.14	3.60	8710.76	4.12
Qwen2.5-VL-7B	8.34	3.21	6209.65	15.83
Qwen2.5-VL-3B	3.79	1.44	8113.90	7.37

Table 12: Efficiency comparison on the Twitter-2015 dataset.

Furthermore, we observe that Qwen2.5-VL-7B exhibits even lower inference latency than Qwen3-VL-2B and Qwen2.5-VL-3B, despite having a larger parameter count. This can be attributed to two factors: first, Qwen2.5-VL-7B has fewer LLM layers (28 vs. 36) (Bai et al., 2025); second, its larger hidden dimension increases the size of per-layer matrix operations, which allows modern GPUs to better saturate memory bandwidth and compute units, thereby achieving higher per-token computation efficiency (higher model bandwidth utilization) (Recasens et al., 2025).