agenticMSA: Agentic Multimodal Sentiment Analysis with Task-Specific and Large Language Model Collaboration

Anonymous ACL submission

Abstract

Multimodal Sentiment Analysis (MSA) faces challenges due to inconsistencies between modalities, such as conflicting sentiment cues from visual, audio, and text data. These modality conflicts make it difficult for previous task-specific samll-scale models to accurately predict sentiment. Although general large multimodal language models (MLLMs) perform well on conflict/hard samples, they can occasionally make errors on simpler samples due to problems like hallucinations or excessive reasoning. To address these issues, we propose agenticMSA, an agentic framework that integrates the strengths of conventional task-specific models and general MLLMs through planning, decision, and reflection agents. The agenticMSA introduces a Modality Conflict Detection (MCD) that identifies modality conflicts, allowing the framework to arrange simpler samples to task-specific models for efficient predictions. For modality conflict samples, we introduce two key modules: 1) Hybrid Collaboration (HC), where decision agents powered by both a task-specific model and a MLLM collaborate to resolve discrepancies. 2) Group Discussion (GD), where multiple MLLM-based decision agents discuss divergent predictions, guided by a reflection agent to reach a consensus. Extensive experiments demonstrate the effectiveness of agenticMSA, achieving state-of-the-art performance on two popular datasets such as CH-SIMS and CMU-MOSI.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to analyze the sentiment from various data types such as vision, audio, language. It has important application value in fields like human-computer interaction and healthcare.

Recently, some progress have been made in exploring better multimodal representation learning for task-specific small-scale models (Zhou et al., 2025; Zhang et al., 2024; Zhao et al., 2024; Yuan et al., 2024; Feng et al., 2024; Zhang et al., 2023). For example, Yuan et al. (2024) introduced a consistency-based pseudo-labeling technique within the semi-supervised framework. Zhao et al. (2024) introduced a sequential multimodal learning strategy that begins by extracting domain-invariant features from textual data and then employs sparsity techniques to identify label-relevant features from video data. Zhang et al. (2023) proposed utilizing the comparatively cleaner language modality to guide the learning of other modalities, thus mitigating the impact of potential noise. However, the performance of these methods remains unsatisfying. As shown in Figure 1, we found this is because conventional task-specific models struggle to achieve better performance when processing modality conflict samples, especially those with significant inconsistencies, a challenge also highlighted in Zhang et al. (2023). In addition to the above studies, the community has an increasing attention on utilizing MLLMs for MSA (Mu et al., 2024; Wang et al., 2024a,c; Lian et al., 2024; Zhang et al., 2025) due to the limitations on the generalization of task-specific small models. However, the

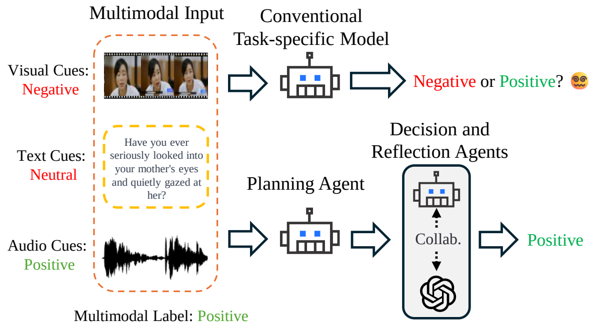


Figure 1: An illustration of modality conflict sample in MSA. Conventional task-specific models often face difficulties in predicting inputs with conflicting sentiment cues across different modalities. In contrast, accurate predictions can be achieved by utilizing a collaborative division of labor among agents.

performance of these methods remains unsatisfactory and requires further exploration due to lack of sufficient research.

Based on our experimental observation (similar to the case shown in Table 9), we find that while MLLMs excel at predicting modality conflict samples (often challenging cases), they may occasionally make errors like hallucinations or overthinking when applied to simpler samples. These issues can result in inaccurate predictions for simpler samples, ultimately affecting the overall performance of the models. Therefore, we suggest leveraging general MLLMs specifically for more complex or challenging samples, which we define as "conflict samples", while relying on task-specific small models for simpler cases. These conflict samples exhibit significant sentiment polarity discrepancies across multiple modalities. For example, a sample where the visual modality conveys a positive sentiment while the audio modality reflects a negative or neutral sentiment would be categorized as a conflict sample. By addressing these conflict samples, we aim to achieve more accurate MSA.

In this paper, we propose agenticMSA, an agentic framework that integrates the strengths of both conventional task-specific models and MLLMs. It introduces a novel way by utilizing planning, decision, and reflection agents to tackle modality conflict samples in MSA, especially through the Group Discussion (GD) to resolve modality conflict samples. Specifically, the Modality Conflict Detection (MCD) module identifies samples with conflicting sentiments, allowing simpler samples to be routed to task-specific models for efficient predictions. For complex conflict samples, the framework uses Hybrid Collaboration (HC), where task-specific models and MLLMs work together to resolve discrepancies. The GD facilitates a reflective dialogue among MLLM-based agents, guided by a reflection agent, to reach consensus when predictions diverge. This enables agenticMSA to tackle both simple and complex cases, achieving state-of-the-art performance on CH-SIMS and CMU-MOSI datasets. In summary, the major contributions can be summarized as:

- We propose an agentic framework (agenticMSA) for MSA that integrates the strengths of both conventional task-specific models and MLLMs.
- We introduce a modality conflict detection mechanism that enables efficient collaboration between MLLM and conventional task-specific models.
- We propose the Group Discussion among agents to make predictions with explanations for modality conflict samples. It facilitate collaborative decision-making among multiple agents to address modality conflict samples.
- The agenticMSA achieves state-of-the-art performance on two popular datasets, including CMU-MOSI and CH-SIMS.

2 Related Work

2.1 Multimodal Sentiment Analysis

Most previous MSA methods (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Hazarika et al., 2020; Zhang et al., 2023; Yu et al., 2021; Zhao et al., 2024; Feng et al., 2024) have focused on learning better unified multimodal representations within the framework of task-specific small-scale models. These efforts have significantly contributed to advancing the field of MSA. For example, Zadeh et al. (2017) utilized the Cartesian product to model relationships between modalities, while Liu et al. (2018) introduced low-rank decomposition to enhance the efficiency of multimodal fusion. With the progress of Transformer architectures, Tsai et al. (2019) introduced Transformers to align long sequences from different modalities effectively. Hazarika et al. (2020) proposed a disentanglement approach to separate modality-invariant and modality-specific representations, making Transformers fuse various modalities from a holistic perspective. Zhang et al. (2023) employed language as a query to guide the learning of other modalities, leading to a more unified multimodal representation. Additionally, Yu et al. (2021) proposed computing pseudo-labels for each modality, facilitating the model to capture both inconsistency and consistency information. It represents a breakthrough among recent non-Transformer-based methods. However, the progress achieved has shown diminishing in recent years. In particular, these methods struggle to process conflict samples where each modality exhibits significant discrepancies in sentiment polarity.

Recently, the MLLMs as mentioned in Section 1, have brought new insights into addressing these challenges. Therefore, we propose the agenticMSA, aiming to provide an effective solution to

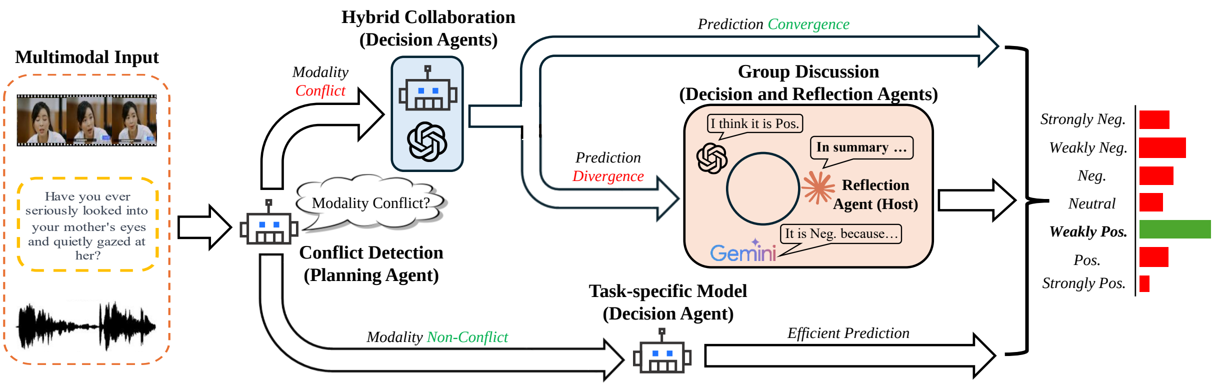


Figure 2: Overall pipeline. The planning agent assigns different streams for processing based on the conflict detection results, and only one of the three streams output the prediction result for a given input.

the issue of tacking conflict samples, thus achieving better MSA.

2.2 Multi-Agent Framework in MLLMs

The multi-agent framework (Qin et al., 2023; Liu et al., 2024; Li et al., 2024; Hong et al., 2024; Chen and Li, 2024; Wang et al., 2024b,d) usually utilizes the collaboration and interaction among multiple specialized agents to solve complex tasks, thus achieving better performance or improved adaptability. Recently, this technique has been explored across various fields such as medical diagnosis and graphical user interface understanding. For example, Li et al. (2024) introduced a framework that integrates diverse open-source medical models to address multiple medical tasks. Wang et al. (2024b) proposed a multi-agent framework including planning, decision, and reflection agents to execute anthropomorphic operations on mobile devices.

While notable progress has been achieved, the application of multi-agent frameworks in MSA remains unexplored. In this paper, we propose the agenticMSA to address modality conflict samples. To the best of our knowledge, this is the first multi-agent-based framework designed for MSA.

2.3 Modality Conflict Detection

A major challenge in MSA is the different modalities (*e.g.,* language, audio, and video) contains inconsistent information, which may mislead the model and degrade its performance. To address this and leverage the strengths of both MLLMs and task-specific models, we propose exploring Multimodal Out-of-Distribution Detection based (OOD-based) techniques, which have recently gained attention in safe machine learning, to detect modality conflict samples. Recent studies have applied OOD Detection to improve performance in multimodal learning (Ming et al., 2022; Wang et al., 2023b; Dong et al., 2024). Inspired by MultiOOD (Dong et al., 2024), we design the MCD module, with further details provided in Section 3.2.

3 Method

3.1 Overview

The agenticMSA framework consists of three key components: modality conflict detection, sentiment prediction for non-conflict samples, and sentiment prediction for modality conflict samples. As shown in Figure 2, given a multimodal input, the modality conflict detection (MCD) module first determines whether the input is a modality conflict sample. For non-conflict samples, an agent powered by a task-specific small-scale model processes the prediction directly and efficiently. For conflict samples, a two-stage process is employed. First, a hybrid collaborative prediction is performed, where decision agents, powered by both MLLMs and task-specific models, collaborate to generate a prediction. If the predictions are diverse, the framework then uses the Group Discussion (GD) module. This module involves multiple MLLMs-powered agents in a reflective discussion to refine the reasoning process and reach a final sentiment decision.

3.2 Modality Conflict

3.2.1 Definition of Modality Conflict

In MSA, each modality contributes unique and complementary information toward determining the overall sentiment. However, inconsistencies can arise when the sentiment polarities expressed across different modalities contradict one another. We define such cases as modality conflict samples.

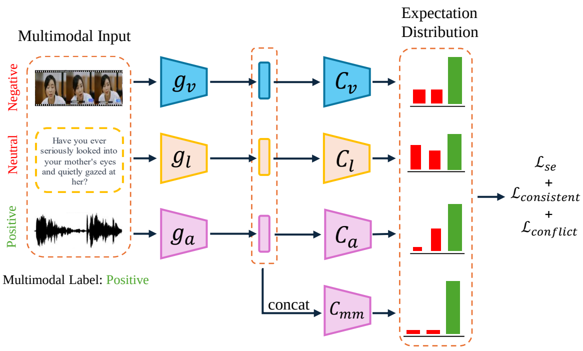


Figure 3: Pipeline of conflict detection agent.

Formally, let X denote the set of multimodal inputs, $\{s_v, s_l, s_a\}$ denote the sentiment polarities from the visual, language, and audio modalities, respectively. Here, $s \in \mathcal{S}$, where $\mathcal{S} = \{\text{negative}, \text{neutral}, \text{positive}\}$ describes the possible sentiment polarities for each modality. For modality conflict samples $X_c \subseteq X$, its sentiment polarities across modalities satisfy:

$$\exists\,(s_i, s_j) \text{ where } s_i \neq s_j, i, j \in \{v, l, a\}, \quad (1)$$

For example, if s_v, s_a, s_l respectively are negative, positive, and positive, the input is classified as X_c due to a conflict between modalities. For non-conflict samples $X_{nc} \subseteq X$, all sentiment polarities across modalities are consistent, *i.e.,* $s_v = s_a = s_l$.

3.2.2 Modality Conflict Detection

With a multimodal input X, the first step of agenticMSA is to determine whether it is a modality conflict sample X_c or a non-conflict sample X_{nc}. To achieve this, we design the MCD module with a planning agent powered by a trained task-specific model as shown in Figure 3. Specifically, modality-specific feature extractors g_m are used to obtain unimodal feature representations $h_m = g(X)$, where $m \in \{v, l, a\}$ and v, l, a represent visual, language, and audio modalities, respectively. These features are then passed into modality-specific classifiers C_m to obtain the class probabilities $p_m = C_m\,(h_m)$, where the probabilities p_m are for the possible sentiment classes: positive, neutral, and negative.

Once the class distributions p_m are obtained, modality conflict detection is performed by computing the mean distance between modalities d_{mm}:

$$d_{mm} = \frac{\sum_{m_i, m_j} \|p_{m_i} - p_{m_j}\|_2}{3}, \quad (2)$$

where $m_i, m_j \in \{v, l, a\}$. Then, a modality conflict sample X_c is identified if the d_{mm} exceeds a threshold $thres$ while the non-conflict sample X_{nc} is identified if the d_{mm} falls below $thres$. The $thres$ are empirically set to 0.65 and 0.77 based on the performance on the validation sets of the CH-SIMS and CMU-MOSI datasets, respectively.

3.2.3 Training Objectives of MCD

The training process of the MCD is guided by $\mathcal{L}_{\text{consistent}}$, $\mathcal{L}_{\text{conflict}}$ and $\mathcal{L}_{\text{se}}$. For non-conflict samples, the probability distributions of any two modalities (denote as m_i and m_j) should be similar for the correct label class y, but dissimilar for all other classes. To enforce this, we minimize the L2 distance of non-label class probabilities:

$$\mathcal{L}_{\text{consistent}} = -\frac{\sum_{m_i, m_j} \|(p_{m_i} - p_{m_j})_{\text{except } y}\|_2}{3}, \quad (3)$$

For modality conflict samples, we maximize the dissimilarity of their entire probability distributions, ensuring no alignment occurs between outputs from conflicting modalities:

$$\mathcal{L}_{\text{conflict}} = -\frac{\sum_{m_i, m_j} \left\|p_{m_i} - p_{m_j}\right\|_2}{3}, \quad (4)$$

In addition, to ensure that both uni-modal and multimodal classifiers (C_{mm}) learn accurate sentiment cues, we apply a cross-entropy loss. The unimodal classifiers optimize as:

$$\mathcal{L}_{\text{se}} = -\frac{\sum_{m'} \sum_{c \in \mathcal{C}} y_c \log(p_{m',c})}{4}, \quad (5)$$

where $p'_m \in \{v, l, a, mm\}$, $c \in \mathcal{C}$ represents the set of multimodal-level sentiment classes, y_c is the one-hot encoded ground truth label for class c, $p_{m',c}$ is the predicted probability for class c from the classifier C'_m.

Overall, the training objective of MCD is:

$$\mathcal{L}_{\text{mcd}} = \alpha(\mathcal{L}_{consistent} + \mathcal{L}_{\text{conflict}}) + \mathcal{L}_{\text{se}}, \quad (6)$$

where α is hyperparameter which is empirically set to 0.5.

3.3 Sentiment Prediction

3.3.1 Prediction for Non-Conflict Samples

For non-conflict samples X_{nc}, a decision agent DATS powered by a task-specific small-scale model is employed to make predictions directly. These simple cases do not require additional reasoning from agents powered by MLLMs, which allows the framework to perform MSA accurately. We denote this process as:

$$\hat{y}_{nc} = \text{DATS}(X_{nc}), \quad (7)$$

In this work, we utilize ALMT (Zhang et al., 2023) with minor changes as the base model.

3.3.2 Prediction for Modality Conflict Samples

For modality conflict samples X_c, a sophisticated reasoning process is employed to ensure accurate predictions. The agenticMSA first utilizes an HC module to generate initial predictions. In this module, two decision agents powered by a task-specific small-scale model and a general MLLM (GPT-4o-mini) are used to analyze X_c. If the predictions of the two agents diverge, the framework transitions to the GD module for further processing.

In the GD, each decision agent GDA_i provides not only its prediction for the given modality conflict sample X_c but also an explanation that supports its decision (The ablation study and more details can be found in Section 4.7 and Appendix B, respectively). If a majority agreement is reached among the agents, the prediction result is directly taken as the majority vote outcome. Otherwise, the final decision requires further discussion and reflection by a reflection agent. It should be noted that a key advantage of GD is the ability to make predictions when the predictions are diverse. Overall, the process can be described as:

$$\hat{y}_c = \begin{cases} \mathrm{HCA}_1(X_c) & \text{if } \mathrm{HCA}_1(X_c) = \mathrm{HCA}_2(X_c) \\ \mathrm{GD}(X_c) & \text{if } \mathrm{HCA}_1(X_c) \neq \mathrm{HCA}_2(X_c) \end{cases} \tag{8}$$

where $\mathrm{HCA}_1(X_c)$ and $\mathrm{HCA}_2(X_c)$ represent the predictions from the agent powered by task-specific small-scale model and the agent powered by MLLM respectively. The prediction process for X_c can be summarized as shown in Algorithm 1.

Algorithm 1: Prediction for X_c

Input: Modality conflict sample X_c
Output: Final sentiment prediction $\hat{y}_c$
Step 1: Hybrid Collaboration (HC):
Initialize decision agents HCA_1 and HCA_2 ;
Obtain prediction: $\hat{y}_{HCA_1} = HCA_1(X_c)$; $\hat{y}_{HCA_2} = HCA_2(X_c)$;
if $\hat{y}_{HCA_1} = \hat{y}_{HCA_2}$ **then**
  Return: $\hat{y}_c = \hat{y}_{HCA_1}$
else
  Step 2: Group Discussion (GD):
  Initiate decision agents: $\{GDA_1, ..., GDA_N\}$;
  Obtain prediction: $\hat{y}_{GDA_i} = GDA_i(X_c)$;
  Count votes for each sentiment class;
  if *majority agreement is achieved* **then**
    Return: $\hat{y}_c$ = Majority Vote Result
  else
    Initiate a reflection agent $GDHA$ as host; Reflection agent obtaion the final prediction: $\hat{y}_c = GDHA(\{\hat{y}_{GDA_1}, ..., \hat{y}_{GDA_N}\})$;
    Return: $\hat{y}_c$

4 Experiments and Analysis

4.1 Dataset

CH-SIMS The dataset (Yu et al., 2020) is a Chinese MSA dataset, containing visual, audio, and language modalities. It comprises 2281 samples in total, divided into 1368 samples for the training set, 456 samples for the validation set, and 457 samples for the test set. Each sample is annotated with a multimodal sentiment score ranging from -1 (negative) to 1 (positive). Moreover, the dataset also provides uni-modal sentiment scores for each modality, with each score in the range of -1 to 1.

CMU-MOSI The dataset (Zadeh et al., 2016) consists of multimodal data, including visual, audio, and language modalities. It comprises a total of 2199 samples, which are divided into 1284 samples for the training set, 229 samples for the validation set, and 686 samples for the test set. Each sample is annotated with a score in multimodal level ranging from -3 (strongly negative) to 3 (strongly positive).

4.2 Evaluation Criteria

We utilized five-class classification accuracy (Acc-5), three-class classification accuracy (Acc-3), binary classification accuracy (Acc-2) and F1 scores as evaluation metrics corresponding to each classification accuracy. For binary classification on the MOSI dataset, previous studies (Yu et al., 2021; Zhang et al., 2023) commonly use two calculation methods: Non-negative/Negative and Negative/Positive. In this work, we use the Non-negative/Negative classification way.

4.3 Implementation Details

Our experiments were conducted on a PC with an NVIDIA A40 GPU. The PyTorch version used is 2.2.1, and the Python version is 3.11.

For multimodal inputs, we used the pre-processed sequences provided by Mao et al. (2022) as inputs for all task-specific models. For MLLMs, video inputs were uniformly sampled into three frames (following Lian et al. (2024)), while audio and language inputs were used without additional processing. In addition, since GPT-4o-mini and Claude-3.5-Sonnet do not support audio inputs, their inputs only consisted of the sampled video frames and language data.

In the MCD module, since the CMU-MOSI dataset does not include uni-modal sentiment annotations, we treated multimodal samples with multimodal-level labels ranging from -1 to 1 as modality conflict samples for training. This choice was motivated by the observation that samples in the range of -1 to 1, representing Weakly Negative, Neutral, and Weakly Positive categories are more prone to exhibiting modality conflict.

The prompt templates are written manually and optimized using GPT-4o-mini (OpenAI, 2023). The prompt templates can be found in Appendix B.

4.4 Baselines

For the task-specific small models, we selected several advanced MSA methods such as TFN (Zadeh et al., 2017), MuLT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), CENET (Wang et al., 2023a), and ALMT (Zhang et al., 2023) for comparison. For MLLMs, due to budget constraints, we only hose all the MLLMs used in our framework, including Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o-mini (OpenAI, 2023) and Gemini-1.5-Flash-8B (Reid et al., 2024), as well as the free call Qwen-VL-Max (Bai et al., 2023) for comparison. In addition, all results were obtained under the same experimental environment to ensure fairness. The results of the task-specific models were reproduced by the authors using the open-source MSA framework MMSA (Mao et al., 2022) with default hyperparameters.

4.5 Performance Comparison

The comparative results are shown in Table 1 and Table 2. For both datasets, agenticMSA consistently achieves the best performance across all metrics compared to task-specific models. For example, agenticMSA achieves a 8.20% relative improvement over the second-best task-specific model (CENET) in terms of Acc-5 on the CH-SIMS dataset. A similar phenomenon can be observed on the CMU-MOSI dataset. Notably, agenticMSA outperforms MLLMs in most metrics, demonstrating its effectiveness. It is worth noting that MLLMs do not perform uniformly well across all evaluation metrics. For example, while GPT-4o-mini achieves the well results for Acc-2 and its corresponding F1, it performs worse on Acc-3. In contrast, agenticMSA shows a more balanced performance across all metrics. It makes agenticMSA more practical and reliable for real-world applications, where consistent results across diverse evaluation critical. Additionally, we have also conducted a comparison with the original results of the task-specific methods, with further details provided in Appendix A.1.

Table 1: Comparison results on CH-SIMS. The best result is highlighted in bold, while the second-best result is indicated with underline .

Method	Acc-5	F1	Acc-3	F1	Acc-2	F1
Qwen-VL-Max	39.61	35.97	50.98	53.35	73.96	67.12
Claude-3.5-Sonnet	43.54	44.36	70.68	69.86	81.18	81.62
GPT-4o-mini	44.63	43.26	66.08	67.79	**82.71**	82.51
Gemini-1.5-Flash-8B	41.79	41.44	70.68	**70.45**	81.62	80.99
TFN	37.94	37.42	65.51	64.39	77.81	77.89
MulT	37.55	37.49	64.60	62.96	77.11	77.13
MISA	36.02	34.07	63.33	61.28	77.38	76.64
Self-MM	42.63	42.92	65.16	64.03	79.90	79.57
CENET	42.67	43.34	64.59	64.66	77.81	77.99
ALMT	40.66	41.02	65.56	64.47	78.08	78.24
agenticMSA	**46.17**	**45.73**	**71.55**	70.26	**82.71**	**82.59**

Table 2: Comparison results on CMU-MOSI. The best result is highlighted in bold, while the second-best result is indicated with underline .

Method	Acc-7	F1	Acc-3	F1	Acc-2	F1
Qwen-VL-Max	42.13	38.83	65.31	68.00	83.38	83.33
Claude-3.5-Sonnet	46.21	44.66	77.70	76.29	85.28	85.28
GPT-4o-mini	39.21	36.94	79.52	**79.67**	85.28	85.27
Gemini-1.5-Flash-8B	46.50	45.35	78.72	77.00	84.55	84.47
TFN	34.08	31.79	66.18	67.08	77.31	78.28
MulT	35.10	32.27	68.86	68.08	78.80	78.89
MISA	42.80	41.43	71.98	73.57	80.67	80.65
Self-MM	46.12	45.21	75.54	75.81	83.51	83.06
CENET	43.82	42.69	72.71	74.10	82.18	82.10
ALMT	43.16	41.94	72.04	73.34	81.50	81.41
agenticMSA	**46.94**	**46.84**	**80.17**	79.49	**86.30**	**86.28**

4.6 Confusion Matrix

Figure 4 presents the confusion matrices for the two datasets. It is evident that the Neutral and Weakly Positive samples exhibit relatively low prediction accuracy on CH-SIMS dataset. This can be attributed to the frequent occurrence of conflicting information across modalities in these samples, making them more challenging to classify. On the CMU-MOSI dataset, the accuracy for samples close to the neutral remains relatively high.

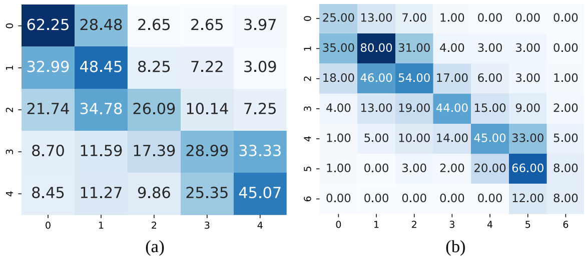


Figure 4: Confusion matrices. (a) CH-SIMS; (b) CMU-MOSI. Note: 0, 1, 2, 3, 4, 5, and 6 represent Strongly Negative, Negative, Weakly Negative, Neutral, Weakly Positive, Positive, and Strongly Positive, respectively.

However, the accuracy for Strongly Negative and Strongly Positive samples is comparatively lower. We attribute this to our way of approximating conflict samples using near-neutral examples (mentioned in Section 4.3), which may overlook the conflicts present in other classes. Therefore, we believe that providing uni-modal labels in future datasets would be beneficial. Addressing the challenge of resolving samples with modality conflicts remains a valuable direction for researching.

4.7 Effect of Each Component

Table 3 presents the effect of removing different agenticMSA components on CH-SIMS datatset, demonstrating the agenticMSA's effectiveness. Some notable points are: 1) When the MCD is removed, we applied GD to all inputs. We can see that there is a significant decline in performance, which is because MLLM-driven agents tend to make errors on simpler samples due to problems like hallucinations or excessive reasoning. This demonstrate that agenticMSA can achieve a balance between performance and efficiency. 2) When the explanation is removed, the decision agent in GD does not need to explain why they make the predictions. In this situation, we use the voting method to determine the final decision and randomly determine the diversified decision results. The performance on all datasets is decreased on most metrics, demonstrating that the explanation is helpful for the agenticMSA in achieving accurate prediction for modality conflict inputs. 3) some differences are observed on the CH-SIMS dataset. Specifically, when the GD module is removed, we make the prediction from the agent powered by the MLLM becomes the final decision, regardless of whether the agents' predictions in the HC module converge or diverge. In this scenario, while Acc-2 and its corresponding F1 score show a slight improvement, Acc-3 experiences a significant drop. This demonstrates agenticMSA's ability to achieve a more balanced performance across all metrics, which is a advantage for real-world applications.

Table 3: Effect of each component.

Method	Acc-5	F1	Acc-3	F1	Acc-2	F1
agenticMSA	**46.17**	45.73	**71.55**	**70.26**	**82.71**	**82.59**
w/o MCD	35.67	35.78	70.46	69.07	81.18	80.84
w/o explanation	45.73	**46.12**	69.58	69.36	80.96	80.69
w/o GD	45.51	44.48	70.68	69.39	80.92	81.09
w/o HC & GD	42.89	42.80	65.65	63.63	78.77	78.47

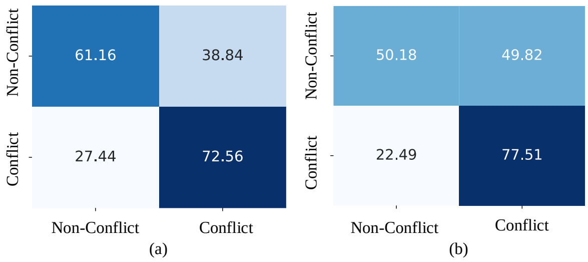


Figure 5: Confusion matrices of modality conflict detection on the CH-SIMS and CMU-MOSI datasets. (a) CH-SIMS; (b) CMU-MOSI.

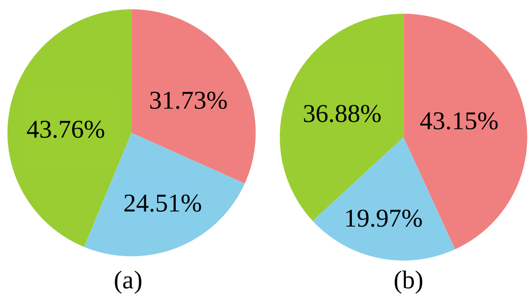


Figure 6: The proportion of data processed by each module. Green denotes the samples processed by agent powered by task-specific small-scale models, blue represents the samples processed by HC, and red corresponds to the samples processed by GD. (a) CH-SIMS; (b) CMU-MOSI.

4.8 Analysis of Modality Conflict Detection

Figure 5 shows the confusion matrix for modality conflict detection. As shown, the detection accuracy for conflict samples exceeds 70% across both datasets, demonstrating the effectiveness of the conflict sample detection. However, it is also evident that there are a higher number of false positives for non-conflict samples in the CMU-MOSI dataset. We attribute this to the lack of uni-modal labels in the MOSI dataset. As mentioned in Section 4.3,

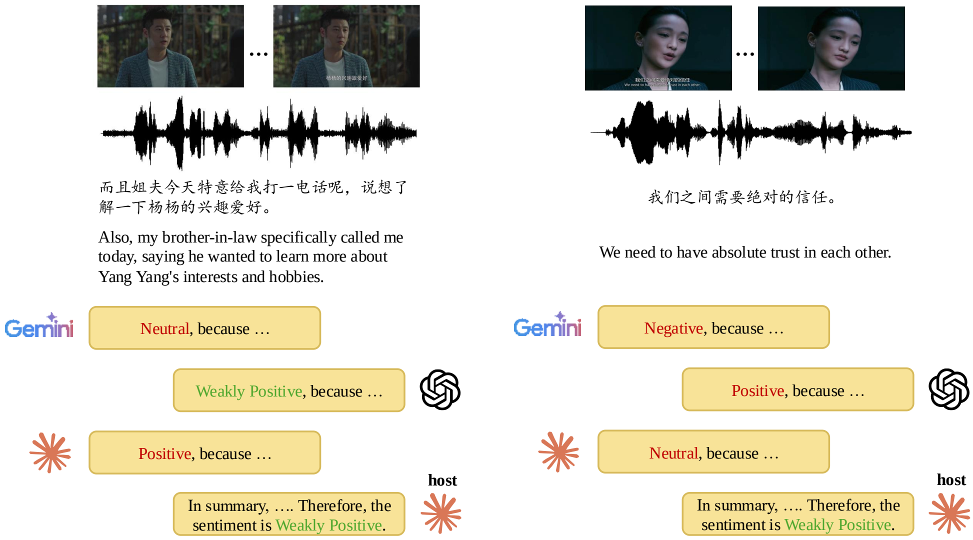


Figure 7: Some examples generated by group discussion. For more detailed examples, please refer to Appendix A.5.

we simplified the classification by treating Weakly Negative, Neutral, and Weakly Positive samples as conflict samples, which may have contributed to the misclassification. In addition, for more analysis of MCD, please refer Appendix A.3.

4.9 Visualization of Module Contribution

Figure 6 illustrates the contributions of different components in the agenticMSA. In Figure 6(a), only 31.73% of the inputs from the CH-SIMS dataset required GD processing. This indicates the MCD can improve the framework's flexibility and achieve a balance between performance and computing resource consumption. Additionally, the HC module proves to be beneficial, as it allows hybrid decisions to be made between MLLMs and task-specific models with small parameters for certain samples, further reducing the resource consumption. A similar phenomenon can be observed on the CMU-MOSI dataset as shown in Figure 6(b). However, it is worth noting that 43.15% of the samples in the CMU-MOSI dataset required GD processing. This was attributed to the inconsistency between the task-specific model and the GPT-4o-mini in the HC stage. Moreover, the absence of uni-modal emotion annotations in the MOSI dataset also had some negative effects on the MCD. Overall, the strength of agenticMSA lies in its ability to achieve a trade-off between performance and efficiency in predictions.

4.10 Case Study

Figure 7 illustrates some examples to demonstrate the effectiveness of agenticMSA, particularly the GD module. Additional detailed cases can be found in Appendix A.5. As shown, in both examples, the decision agents express differing opinions during the discussion. However, the host (reflection agent) consistently makes correct predictions by reflecting on and summarizing the opinions. This demonstrates the advantage of the GD mechanism over traditional voting strategies and underscores the overall effectiveness of our framework. Furthermore, we utilize Claude-3.5-Sonnet as the base model for the reflection agent. For ablation experiments with different models as the reflection agent, please refer to Appendix A.4.

5 Conclusion

In this paper, we introduce agenticMSA, a agentic framework for Multimodal Sentiment Analysis (MSA) that addresses the challenges of modality conflicts. Extensive experiments on the CH-SIMS and CMU-MOSI datasets demonstrate the framework's superior performance, achieved through the integration of task-specific models, Hybrid Collaboration (HC), and Group Discussion (GD). The framework's ability to dynamically allocate resources and combine MLLMs with conventional task-specific models achieves a trade-off between performance and efficiency in predictions. The state-of-the-art performance highlight the potential of agenticMSA to advance the field of MSA, offering a new solution to the modality inconsistence.

Limitations

Despite the encouraging results, agenticMSA still has its limitations. Modality conflict samples often exhibit conflicting sentiment tendencies across different modalities. In MSA datasets, only a few contain uni-modal sentiment annotations, which limits the capability of modality conflict detection and consequently limits the universality and overall performance of the framework.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, AI. Technical report.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Wei Chen and Zhiyuan Li. 2024. Octopus v3: Technical report for on-device sub-billion multimodal AI agent. *CoRR*, abs/2404.11459.

Hao Dong, Yue Zhao, Eleni Chatzi, and Olga Fink. 2024. MultiOOD: Scaling out-of-distribution detection for multiple modalities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. 2024. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14755–14766, Miami, Florida, USA. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 1122–1131. ACM.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogagent: A visual language model for GUI agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 14281–14290. IEEE.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. GPT-4V with emotion: A zero-shot benchmark for generalized emotion recognition. *Inf. Fusion*, 108:102367.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2024. Llava-plus: Learning to use tools for creating multimodal agents. In *Computer Vision - ECCV 2024 - 18th European Conference*, volume 15105 of *Lecture Notes in Computer Science*, pages 126–142. Springer.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher 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 (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 204–213, Dublin, Ireland. Association for Computational Linguistics.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. 2022. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Jie Mu, Wei Wang, Wenqi Liu, Tiantian Yan, and Guanglu Wang. 2024. Multimodal large language model with lora fine-tuning for multimodal sentiment analysis. *ACM Trans. Intell. Syst. Technol.*

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models. *CoRR*, abs/2304.08354.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2023a. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Trans. Multim.*, 25:4909–4921.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. 2023b. CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 1802–1812. IEEE.

Jie Wang, Yichen Wang, Zhilin Zhang, Jianhao Zeng, Kaidi Wang, and Zhiyang Chen. 2024a. Sentixrl: An advanced large language model framework for multilingual fine-grained emotion classification in complex text environment. *CoRR*, abs/2411.18162.

Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024b. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *CoRR*, abs/2406.01014.

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

Yiying Wang, Xiaojing Li, Binzhu Wang, Yueyang Zhou, Han Ji, Hong Chen, Jinshi Zhang, Fei Yu, Zewei Zhao, Song Jin, Renji Gong, and Wanqing Xu. 2024d. Peer: Expertizing domain-specific tasks with a multi-agent framework and tuning methods. *Preprint*, arXiv:2407.06985.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 10790–10797. AAAI Press.

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

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

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.*, 31(6):82–88.

Haoyu Zhang, Xiaoying Tang, Wei Liu, Jian Luan, and Tianshu Yu. 2025. Prompt-guided distillation from multimodal large language models to task-specific models for multimodal sentiment analysis.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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, Singapore. Association for Computational Linguistics.

Xianbing Zhao, Lizhen Qu, Tao Feng, Jianfei Cai, and Buzhou Tang. 2024. Learning in order! A sequential strategy to learn invariant features for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024*, pages 9729–9738. ACM.

Ying Zhou, Xuefeng Liang, Han Chen, Yin Zhao, Xin Chen, and Lida Yu. 2025. Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, 114:102663.

A Supplementary Analysis

A.1 Additional Comparison

As shown in Table 4 and Table 5, although there might be some differences in experimental environments and settings, we have still included a comparison with the original results of additional methods to emphasize the effectiveness of our framework. These methods include TFN (Zadeh et al., 2017), MuLT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), ALMT (Zhang et al., 2023), and KuDA (Feng et al., 2024). Similiar to Section 4.5, the results also show that agenticMSA can achieves leading performance, demonstrating the superiority of the framework.

Table 4: Additional comparison results on CH-SIMS. The best result is highlighted in bold. a represents the result is from (Yu et al., 2020)

Method	Acc-5	F1	Acc-3	F1	Acc-2	F1
TFN[a]	38.38	-	64.46	-	80.66	81.62
MulT[a]	35.34	-	65.03	-	77.94	79.10
ALMT	45.73	-	68.93	-	81.19	81.57
KuDA	43.54	-	66.52	-	80.74	80.71
agenticMSA	**46.17**	**45.73**	**71.55**	**70.26**	**82.71**	**82.59**

Table 5: Additional comparison results on CMU-MOSI. The best result is highlighted in bold.

Method	Acc-7	F1	Acc-3	F1	Acc-2	F1
TFN	28.7	-	-	-	73.9	73.4
MISA	42.3	-	-	-	81.8	81.7
Self-MM	-	-	-	-	84.00	84.42
MMIM	46.65	-	-	-	84.14	84.00
ALMT	**49.42**	-	-	-	84.55	84.57
KuDA	47.08	-	-	-	84.40	84.48
agenticMSA	46.94	**46.84**	**80.17**	**79.49**	**86.30**	**86.28**

A.2 Additional Analysis of Each Component

Table 6 presents the performance of agenticMSA with different components removed on the CMU-MOSI dataset. Similar to the results on the CH-SIMS dataset (as shown in Section 4.7), agenticMSA demonstrates strong performance across most metrics, with performance drops when any components are removed. This consistent trend across multiple datasets further validates the effectiveness of our framework.

Table 6: Effect of each component on the CMU-MOSI dataset.

Method	Acc-7	F1	Acc-3	F1	Acc-2	F1
agenticMSA	46.94	46.84	**80.17**	**79.49**	**86.30**	**86.28**
w/o MCD	38.76	38.92	76.38	74.49	82.07	81.86
w/o explanation	47.08	47.04	79.88	78.84	84.84	84.81
w/o GD	**48.54**	**47.35**	75.80	76.60	83.38	83.37
w/o HC & GD	47.52	47.24	77.55	77.31	84.11	84.10

A.3 Additional Analysis of Regularization in MCD

Figure 8 illustrates the distribution of the mean distance between modalities used for modality conflict detection, both with and without the regularization terms $\mathcal{L}consistant$ and $\mathcal{L}conflict$, applied during training on the CH-SIMS and CMU-MOSI test datasets. As shown in Figure 8(a) and Figure 8(c), when regularization is included in the training process, the modality distance distribution is distinguishable, facilitating the identification of modality conflict samples. In contrast (Figure 8(b) and Figure 8(d)), when regularization is removed, the distances between the modalities of most samples are concentrated, making it more difficult to detect modality conflict samples. This demonstrate the effectiveness of regularization in improving the performance of MCD.

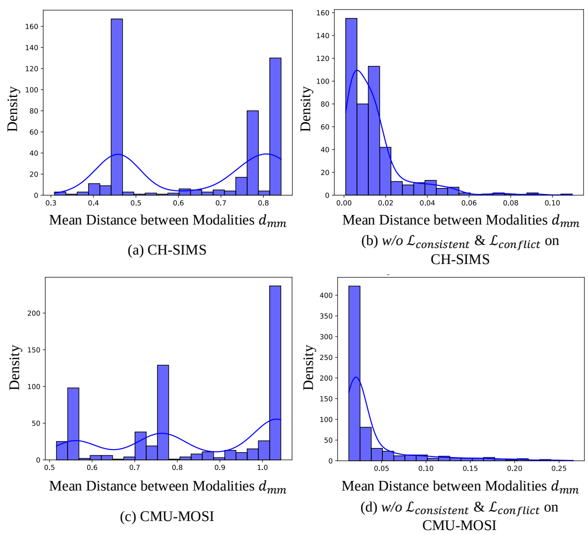


Figure 8: Visualization of the mean distance distribution used for modality conflict detection, with and without $\mathcal{L}_{\text{consistant}}$ and $\mathcal{L}_{\text{conflict}}$, using Kernel Density Estimation.

A.4 Additional Analysis of Reflection Agent Model Selection

Table 7 presents the performance results when different models are used as reflection agents in the group discussion. The results demonstrate that both Claude-3.5-Sonnet and GPT-4o-mini exhibit well-balanced performance across both datasets. In contrast, when the reflection agent is powered by Gemini-1.5-Flash-8B, performance decreases across all datasets. This suggests that tasks involving summarization and reflection benefit from larger models, which are better suited for these more complex tasks.

Table 7: Performance when reflection agents use different models.

Method	CH-SIMS					
	Acc-5	F1	Acc-3	F1	Acc-2	F1
Claude-3.5-Sonnet	**46.17**	**45.76**	71.54	**70.26**	**82.71**	**82.59**
Gemini-1.5-Flash-8B	**46.17**	45.60	71.33	70.06	82.28	82.02
GPT-4o-mini	45.95	45.56	**71.55**	70.23	82.49	82.46
Method	**CMU-MOSI**					
	Acc-7	F1	Acc-3	F1	Acc-2	F1
Claude-3.5-Sonnet	46.94	46.84	**80.17**	**79.49**	86.30	86.28
Gemini-1.5-Flash-8B	46.50	46.49	78.43	78.03	84.55	84.54
GPT-4o-mini	**47.67**	**47.64**	**80.17**	79.48	**86.88**	**86.87**

A.5 Supplement of Case Study

As shown in Table 8 and Table 9, we present the details of the Group Discussion, including the prediction with explanation of each agent, when processing modality conflict inputs. It can be observed that even when there are disagreements among the decision agents, the reflective agent ultimately makes the correct decision through reflection and summarization.

B Prompt Template

Table 10 shows the prompt template used in agenticMSA for reference.

Table 8: An example generated by group discussion of agenticMSA. The results from the decision agent powered by the task-specific model are provided for demonstration purposes and are not part of the discussion.

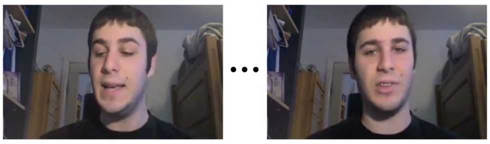
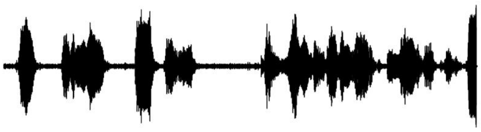
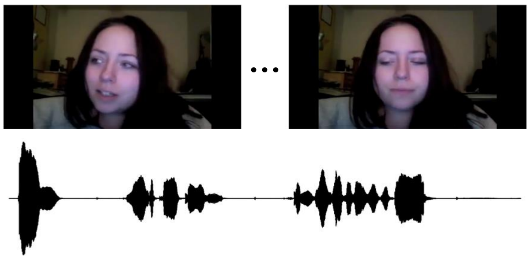
Input	
	the only actor who can really sell their lines is ERIN ECKART
	Label: Weakly Negative
Agent	**Output**
Decision agent (Task-specific, ALMT)	Negative
Decision agent (Gemini-1.5-Flash-8B)	**Analysis**: The video shows a young man speaking. His facial expressions are neutral, neither overly positive nor negative. There's no discernible emotional tone in his body language or posture. The audio is a statement, not an enthusiastic declaration or a sarcastic remark. The text "THE ONLY ACTOR WHO CAN REALLY SELL THEIR LINES IS ERIN ECKART" is a factual claim, not a strong expression of praise or condemnation. While the speaker is making a definitive statement, it isn't laden with emotional intensity. There is no indication of strong emotion behind the claim; it appears to be simply an opinion. **Prediction**: Neutral
Decision agent (GPT-4o-mini)	**### Analysis ###** **Analysis**: The text segment states, "THE ONLY ACTOR WHO CAN REALLY SELL THEIR LINES IS ERIN ECKART." The phrase "THE ONLY ACTOR" suggests exclusivity and a high level of praise directed towards Erin Eckart, indicating a strong positive sentiment. Describing Erin as someone who can "really sell their lines" reinforces this positive sentiment, implying efficacy and excellence in performance. The use of "really" further strengthens the sentiment, suggesting that the speaker feels strongly about the capability in question. Without additional non-verbal cues to consider from the video, the verbal content alone is compelling enough to support a positive interpretation. **Prediction**: Positive

Table 8: An example generated by group discussion of agenticMSA. (continuation of Table)

Agent	Output
Decision agent (Claude-3.5-Sonnet)	**Analysis**: In this video sequence, I observe someone speaking with a relatively flat and unenthusiastic facial expression across all three frames. The speaker appears to be delivering criticism about acting performances, specifically singling out one performer (Erin Eckart) as the only one who can "really sell their lines." The visual cues show a rather serious demeanor, with minimal emotional expression or animation. Combined with the text, which implies that other actors were not as capable, this creates an overall negative sentiment. While there is one positive note about Erin Eckart's performance, the statement is framed in a way that criticizes everyone else by comparison. **Prediction**: Negative. The combination of the speaker's stern expression and the critical nature of the statement, which essentially dismisses the performances of all other actors except one, justifies a negative sentiment classification. While not harshly negative enough to be "strongly negative," it still expresses clear disapproval of most of the acting performances in question.
Reflection agent (host, Claude-3.5-Sonnet)	**Output**: Let's analyze this multimodal content carefully: 1. Text Analysis: - The text is a critical commentary about acting performances - Uses ""ONLY"" which is exclusionary and dismissive of other actors - While praising one actor (Erin Eckart), it implicitly criticizes all others 2. Visual Analysis: - The video shows a straightforward, direct delivery - The demeanor appears matter-of-fact rather than overtly negative or positive - The expression conveys seriousness while delivering the critique 3. Model Predictions: - Model A (Positive): This seems least accurate given the critical nature - Model B (Neutral): Has merit due to the measured delivery style - Model C (Negative): Aligns well with the content's critical nature 4. Synthesis: - The text is clearly critical - The delivery style is relatively measured - While praising one actor, the statement inherently criticizes everyone else - The combination suggests disapproval, though not extremely harsh The most reasonable prediction appears to be Model C's negative assessment, though I would qualify it as ""weakly negative"" rather than strongly negative because: - While critical, it's not an outright condemnation - There is one positive element (praise for Erin Eckart) - The delivery is measured rather than angry or hostile **Prediction**: Weakly Negative

Table 9: An example generated by group discussion of agenticMSA. The results from the decision agent powered by the task-specific model are provided for demonstration purposes and are not part of the discussion.

Input



um yeah the characters didn't care whether they lived or died

Label: Negative

Agent	**Output**
Decision agent (Task-specific, ALMT)	Negative
Decision agent (Gemini-1.5-Flash-8B)	**Analysis**: The video shows a woman speaking. The audio contains a hesitant start ("um"), followed by a statement about characters in a story not caring about living or dying. The tone of voice in the audio is neutral; there's no discernible emotionality, nor any vocal emphasis suggesting strong feelings about the statement. The facial expressions in the video frames are also neutral. There's no visible sadness, anger, joy, or other strong emotions. The statement itself, while potentially commenting on a story's lack of emotional depth, doesn't inherently express a positive or negative sentiment from the speaker's perspective. The speaker's words describe a characteristic of the story, rather than a personal evaluation of it. **Prediction**: Neutral
Decision agent (GPT-4o-mini)	**### Analysis ###** The text states a rather cynical view about characters not caring whether they lived or died. The phrase "DIDNT CARE" suggests a sense of apathy or indifference towards the characters' fates, which reflects negatively on the narrative or their development. The use of all capital letters emphasizes this sentiment, indicating strong feelings about the matter. While the speaker might not explicitly convey deep anger or frustration, the overall tone of the text carries a strongly negative attitude, suggesting discontent with how the characters were portrayed. The absence of concern from the characters implies a lack of meaningful stakes in the story, which could suggest that the viewer is dissatisfied with the narrative quality. Without seeing the facial expressions or gestures in the video, it is difficult to determine the complete sentiment as presented in the speaker's delivery. However, given the sentiment of the text and the likely emotional weight behind it, one can infer that the speaker's overall tone is negative. **### Prediction ###** Strongly Negative

Table 9: An example generated by group discussion of agenticMSA. (continuation of Table)

Agent	Output
Decision agent (Claude-3.5-Sonnet)	**Analysis**: The video frames show someone with a rather solemn and unenthusiastic expression throughout. Their facial features remain consistently downcast, with lowered eyes and a serious demeanor that suggests disappointment or criticism. The text is written in all caps, which can denote emphasis or frustration, and discusses characters being indifferent about mortality ("DIDNT CARE WEATHER THEY LIVED OR DIED"), which is a notably dark and negative observation. The speaker appears to be criticizing this character trait with disapproval. The combination of the dejected facial expressions and the critical content about characters' apathy toward life and death creates a clear negative sentiment. **Prediction**: Negative
Reflection agent (host, Claude-3.5-Sonnet)	**Analysis**: Looking at all available information and model predictions, let me analyze each component: 1. Text Analysis: - Written in all caps, suggesting emphasis or intensity - Content expresses criticism about characters' lack of self-preservation instinct - Demonstrates viewer frustration with character development/writing 2. Visual Analysis: - The images show consistent and relatively subdued expressions - The demeanor appears somewhat flat or unenthusiastic - The visual cues align with someone expressing disappointment or criticism 3. Model Predictions: Model A (Strongly negative): - Makes strong arguments about the cynical nature of the comment - Correctly identifies the criticism in the content - May be slightly overestimating the intensity Model B (Neutral): - Seems to underestimate the clear negative sentiment in both text and visuals - Doesn't adequately account for the criticism being expressed - Likely too conservative in its assessment Model C (Negative): - Provides a balanced interpretation - Acknowledges the negative sentiment without overstating it - Best aligns with both textual and visual evidence After weighing all factors, Model C's "negative" prediction appears most accurate. The comment clearly expresses disappointment with character writing, but the visual delivery suggests regular criticism rather than intense negativity. While Model A's "strongly negative" might be too extreme and Model B's "neutral" too mild, a simple "negative" rating best captures the overall sentiment. **Prediction**: Negative

Table 10: Prompt template. "Candidate Labels" represent the Candidate Labels of each dataset.

Agent	Prompt
Decision agent with video-audio-language inputs	**### Background ###** Assume that you are a multimodal sentiment analysis expert. We provide you with a video-audio-text pair, where the video contains 3 temporally uniformly sampled frames with the speaker's content. The available video-audio-text pairs is: **### Video-audio-text Pair ###** Video: Please refer to the video input. Audio: Please refer to the audio input. Text: text **### Response Requirements ###** Please ignore the speaker's identity and concentrate on the sentiment cues. Analyze the provided video-audio-text pair and give a prediction. The video-audio-text pair should be classified as one of the sentiment following: {Candidate Labels}. **### Output Format ###** Analysis: Provide a detailed analysis of the sentiment cues in the video-audio-text pair. Prediction: Directly provide a class from {Candidate Labels}.
Decision agent with video-language inputs	**### Background ###** Assume that you are a multimodal sentiment analysis expert. We provide you with a video-text pair, where the video contains 3 temporally uniformly sampled frames with the speaker's content. The available video-text pairs is: **### Video-text Pair ###** Video: Please refer to the video input. Text: text **### Response Requirements ###** Please ignore the speaker's identity and concentrate on the sentiment cues. Analyze the provided video-text pair and give a prediction. The video-text pair should be classified as one of the sentiment following: {Candidate Labels}. **### Output Format ###** Analysis: Provide a detailed analysis of the sentiment cues in the video-text pair. Prediction: Directly provide a class from {Candidate Labels}.

Table 10: Prompt template. ”Candidate Labels“ represent the Candidate Labels of each dataset. (continuation of Table)

Agent	Prompt
Reflection agent in group discussion	**### Background ###** Assume that you are a multimodal sentiment analysis expert. There is a video-audio-text pair, along with the analysis process and prediction results from multiple models. You need to read it carefully and analyze it comprehensively. **### Video-audio-text Pair ###** Video: Please refer to the video input. Audio: Please refer to the models' analysis. Text: text **### Prediction of Model A ###** {gpt4omini_preds} **### Prediction of Model B ###** {gemini_1_5_flash_8b_preds} **### Prediction of Model C ###** {claude_preds} **### Task ###** Analyze the video-audio-text pair and the prediction results of all models. Give a most reasonable prediction refer to the three models and provide the reason for your prediction. You need to: 1. Evaluate the reasons for each model, indicating which are strong supporting arguments and which may be misleading. 2. Weigh opinions between different models, especially when they disagree. 3. Check for any conflicts or inconsistencies and try to resolve them. 4. Combine all available information to generate a final forecast and provide a detailed explanation of why this particular outcome was chosen. **### Output Format ###** Analysis: Provides analysis of the above tasks. Prediction: Provide a final decision from {Candidate Labels}.

C Impact Statement

This paper presents work whose goal is to advance the field of Multimodal Sentiment Analysis and Multimodal Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.