Mutual Refinement Distillation for Multimodal Emotion Recognition: Interactive Learning and Reverse Curriculum for Complex Sample Classification

Liu Linsong , Yu Gu * , He Zhang , Shuang Wang , Chenyu Li , Quan Ande , Haixiang Lin

Posted Date: 29 January 2026

doi: 10.20944/preprints202601.2321.v1

Keywords: mutual refinement distillation; distillation; modal interaction calibration; interactive learning constraints; reverse curriculum learning

 doi:10.20944/preprints202601.2321.v1

Article

Mutual Refinement Distillation for Multimodal Emotion Recognition: Interactive Learning and Reverse Curriculum for Complex Sample Classification

Linsong Liu [1], Yu Gu [1,*], He Zhang [2], Shuang Wang [1], Chenyu Li [1], Quan Ande [1] and Haixiang Lin [3]

[1] School of Artificial Intelligence, Xidian University, Xi'an, China
[2] School of Journalism and Communication, Northwest University, Xi'an, China
[3] Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
* Correspondence: guyu@xidian.edu.cn; Tel.: +86-1780-924-0689

Abstract

With the rapid advancement of speech emotion recognition, the transition from unimodal to multimodal approaches has become inevitable. However, multimodal methods introduce new challenges, particularly classification ambiguity in complex samples when compared to unimodal approaches. To address this, we propose a Mutual Refinement Distillation (MRD) method, which incorporates three key components: (1) Modal Interaction Calibration, enhancing classification accuracy for complex samples; (2) Interactive Learning Constraints, mitigating overfitting; and (3) Reverse Curriculum Learning, further improving model robustness. Experiments with the MELD and IEMOCAP datasets demonstrate that our approach outperforms state-of-the-art methods in emotion recognition, achieving a notable 6.07% improvement over the baseline on IEMOCAP.

Keywords: mutual refinement distillation; distillation; modal interaction calibration; interactive learning constraints; reverse curriculum learning

1. Introduction

Traditional speech emotion recognition (SER) approaches, which primarily rely on acoustic features, struggle to capture the complexity and nuance of real-world emotional expressions [1,2]. Recent advancements in deep learning, including Convolutional Neural Networks (CNNs), Long Short Term Memory (LSTM) networks [2,3], and attention mechanisms [4], have significantly improved SER performance. However, even with these methods, the limitation of single modality lies in the lack of additional auxiliary information, which can easily lead to information bottlenecks and prevent significant performance improvement. Unlike unimodal approaches, integrating multiple modalities such as facial expressions, physiological signals, and textual content has led to a paradigm shift, enabling more accurate and contextually rich emotion recognition [5,6]. Common multimodal fusion techniques include early fusion, late fusion, and hybrid fusion methods [7].

Multimodal methods primarily aim to integrate information from multiple modalities to enhance speech emotion recognition [8]. However, different fusion strategies may introduce interference factors, leading to classification ambiguity. In some cases, this interference can cause a multimodal model to perform worse than a unimodal one for certain samples. Despite this issue, previous research has not explicitly addressed it, as most studies focus on adjusting modality contributions and filtering out interfering information [9–11], both of which are inherently efforts to reduce ambiguity.

We conducted pilot experiments to investigate classification ambiguity in multimodal methods, specifically using the state-of-the-art CMERC multimodal approach [12]. As shown in Table 1 and Table 2, we define Multimodal Correct but Text Incorrect (MCTI) as instances where the multimodal model (integrating text, audio, and visual features) predicts correctly while the text-only model fails.

 doi:10.20944/preprints202601.2321.v1

Conversely, Multimodal Incorrect but Text Correct (MITC) refers to cases where the multimodal model is incorrect, but the text-only model succeeds.

Given that the most significant ambiguity occurs between the text modality and the multimodal model, with closely matched classification accuracy, we focused on comparing these two models and used them in subsequent experiments. We also compared other modalities and found that the classification ambiguity among them was not as pronounced as that between the text and multimodal models, so we did not conduct additional experiments with the other modalities. This experiment demonstrates the classification ambiguity between text modality and multimodal models, particularly evident on the IEMOCAP dataset.

Table 1. IEMOCAP dataset performance on baseline model by emotion category

Datasets	Happy	Sad	Neutral	Angry	Excited	Frustrated	Total
MCTI	22	19	46	11	92	48	238
MITC	24	16	15	14	20	23	112
Total Data	144	245	384	170	299	381	1623

Table 2. MELD dataset performance on baseline model by emotion category

Datasets	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Total
MCTI	60	7	0	1	1	1	5	75
MITC	14	8	6	27	35	3	31	124
Total Data	1256	281	50	208	402	68	345	2610

Based on the experimental results, we observe that unimodal and multimodal approaches often exhibit classification ambiguity in certain specific samples within multimodal SER. We define these samples as complex samples. This phenomenon is further supported by prior research, such as CMERC, which demonstrates that removing certain modalities can, in some cases, lead to increased classification confidence [12].

Previous studies have attempted modal weighted fusion to address classification ambiguity, employing methods such as transformers [9,13] and their variants, including EmoCaps, MPT-HCL and CFN-ESA [14–16]. Additionally, Graph Neural Networks (GNNs) have been utilized in models like [17], such as M3Net, MGLRA and CMERC [12,18,19]. Beyond optimizing modal fusion, other approaches focus on eliminating interference to reduce ambiguity. For instance, feature decoupling techniques identify shared and modality-specific features to extract only the most useful information [10,20–23]. Alternatively, knowledge distillation leverages stronger modalities to guide weaker ones, preventing interference from weaker modes [24]. Although these methods have made significant progress in reducing ambiguity, even the most advanced models struggle to eliminate it entirely.

To address this issue, we propose Mutual Refinement Distillation (MRD), a method designed to reduce classification ambiguity as much as possible in both multimodal and unimodal models without altering their core frameworks. As illustrated in the pipeline of Figure 1, MRD consists of three key components: **(1) Modal Interaction Calibration (MIC):** Inspired by TelME [25], we propose mutual distillation (MD) as an alternative to traditional distillation. MD enables knowledge transfer between multimodal and unimodal models to address classification ambiguity in complex samples. Unlike standard distillation, where one model is guided by another, MD allows both models to learn from each other equally, offering a balanced approach to resolving ambiguity [26]. **(2) Interactive learning constraints (ILC):** To prevent the models from overfitting to each other's characteristics during mutual distillation, we introduce two refactoring losses. These losses ensure that the models retain their distinct characteristics while still benefiting from mutual learning [27]. **(3) Reverse Curriculum Learning (RCL):** To enhance the model's focus on complex samples, we introduce reverse curriculum learning (RCL). Unlike traditional curriculum learning [28] that begins with simpler samples, RCL starts with complex samples that exhibit higher classification ambiguity and then moves to simpler ones.

 doi:10.20944/preprints202601.2321.v1

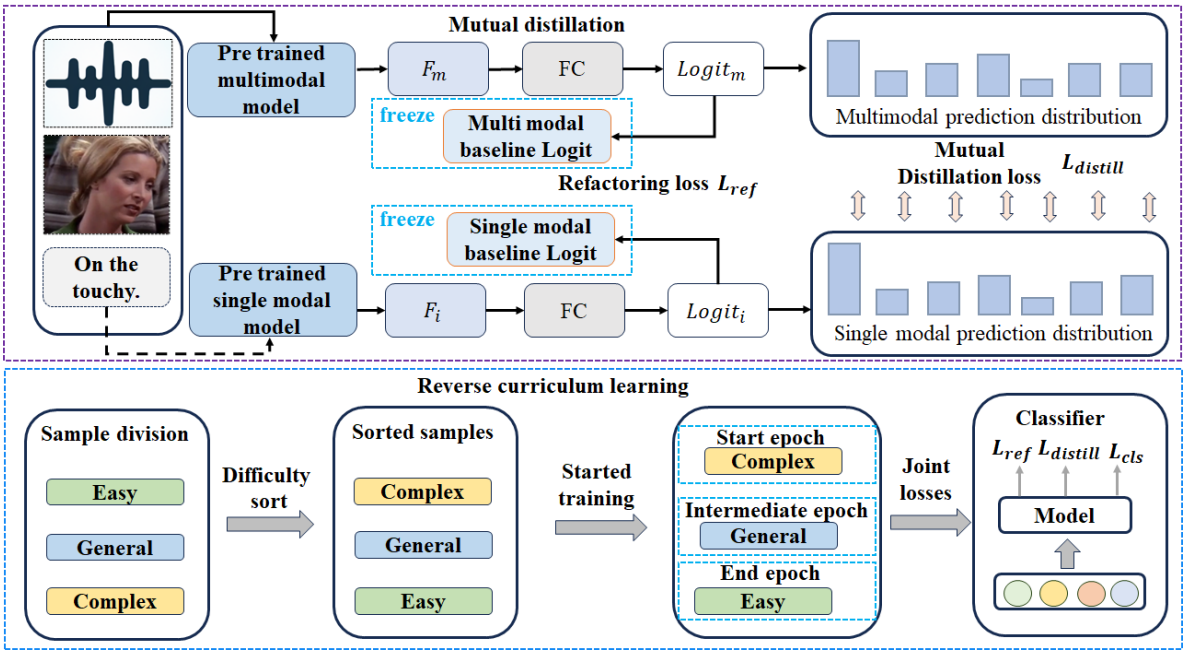

Figure 1. The pipline of Mutual Refinement Distillation. Mathematical symbols are consistent with the formulas in the paper.

2. Materials and Methods

For textual feature extraction, we apply layer normalization and average the final four hidden layers of the RoBERTa model [29], following the approach of G. Tu et al. [12]. For audio features, we use OpenSMILE [30], while for visual features, we utilize a pre-trained DenseNet model [31]. Additionally, we employ the CMERC method to obtain pretrained multimodal (text, audio, and visual) and unimodal (text-only) models [12].

2.1. Modal Interaction Calibration: Interaction Calibration Between Unimodal and Multimodal Models

To enhance the effectiveness of both text-only and multimodal models, we employ mutual distillation, which calibrates their respective useful information. The core of this method lies in defining appropriate model outputs and loss functions. We first obtain a pre-trained multimodal model M_{multi} and a pre-trained text-only model M_{text}. The output of the multimodal model can be represented as:

$$Y_{multi} = M_{multi}(X_{multi}) \tag{1}$$

Here, X_{multi} represents the primary multimodal features. Similarly, the output of the text modality model is defined as:

$$Y_{text} = M_{text}(X_{text}) \tag{2}$$

where X_{text} represents the primary features of the text modality. To achieve MIC, we define two distillation losses(DL), which capture the bidirectional knowledge transfer: (1) the process of learning the multimodal model from the text modality model, and (2) the process of learning the text modality model from the multimodal model.

First, the distillation loss for the multimodal model is defined as:

$$L_{distill\_multi} = -\sum_i Y_{text}(i) \log\left(\frac{e^{Y_{multi}(i)/T}}{\sum_j e^{Y_{multi}(j)/T}}\right) \tag{3}$$

where T is the temperature parameter, which smooths the output probability distribution to facilitate learning. In the above formula, i and j denote the indices of the respective samples, a convention that applies to the following equations as well.

 doi:10.20944/preprints202601.2321.v1

Next, the distillation loss for the text modality model is defined as:

$$L_{distill\_text} = -\sum_i Y_{multi}(i) \log\left(\frac{e^{Y_{text}(i)/T}}{\sum_j e^{Y_{text}(j)/T}}\right) \tag{4}$$

Based on this, we formulate the total distillation loss function by weighting and summing the two DL:

$$L_{distill} = \alpha L_{distill\_multi} + \beta L_{distill\_text} \tag{5}$$

In this formula, α and β are weight coefficients that balance the contribution of each DL.

2.2. Interactive Learning Constraints: Preventing Overfitting in Single-Modality and Multimodal Models

To prevent the two models from overfitting and losing their unique characteristics during interactive learning, we introduce two refactoring losses(RL) as constraints.

First, we define the RL, which quantifies the difference between each model's output before and after distillation. Specifically, we freeze the outputs of the two pre-trained models before distillation, denoted as Y_{frozen_multi} and Y_{frozen_text}.

The refactoring loss for the multimodal model, L_{ref_multi}, is formulated as:

$$L_{ref\_multi} = \frac{1}{n}\sum_i \left|Y_{multi}(i) - Y_{frozen\_multi}(i)\right|^2 \tag{6}$$

Similarly, the RL for the text modality model, denoted as L_{ref_text}, is formulated as:

$$L_{ref\_text} = \frac{1}{n}\sum_i \left|Y_{text}(i) - Y_{frozen\_single}(i)\right|^2 \tag{7}$$

where n represents the number of samples, and $|\cdot|^2$ denotes the squared difference.

Next, we combine the RL for the multimodal and text-modality models to formulate the total RL:

$$L_{ref} = \gamma L_{ref\_multi} + \delta L_{ref\_single} \tag{8}$$

In this equation, γ and δ are weight coefficients for RL, used to balance the contributions of different RL components.

This loss helps the text modality model retain its core features while benefiting from knowledge distillation, preventing excessive adaptation to multimodal influences. For the multimodal distillation constraint, it ensures balanced knowledge integration across modalities, avoiding overfitting to text modality.

2.3. Reverse Curriculum Learning: Prioritizing Complex Samples

In recent years, Curriculum Learning [28] has proven effective in enhancing the training efficiency and generalization ability of deep models by initially focusing on simple samples and gradually introducing more complex ones during training. However, this approach assumes that the model starts with little understanding of the task and needs to build knowledge incrementally. In the context of Knowledge Distillation, student models often already possess some level of representational and cognitive abilities, particularly when teacher models offer strong distillation performance or when pre-trained models are used. If an "easy-first, difficult-later" strategy is still applied, the model may quickly converge on simple samples, limiting the time and resources available to tackle complex samples, which can consequently restrict the model's performance improvement on challenging tasks [32].

In contrast, recent approaches such as Reverse Curriculum Learning [33] and Hard Example Mining [34] suggest that training should focus more on complex samples that the model has not yet mastered. Reverse Curriculum Learning stimulates the model's ability to learn high-information

 doi:10.20944/preprints202601.2321.v1

features by prioritizing challenging samples, effectively focusing training resources on the model's weaknesses and promoting further learning capability improvement. In knowledge distillation, this strategy aligns better with the needs of student models: they can already handle simple samples well, and greater improvements can be achieved through further learning and fitting of complex misclassified samples.

To further enhance the performance of our MIC method, we incorporate the concept of Reverse Curriculum Learning into the knowledge distillation process. This involves prioritizing the training of complex samples that are challenging for both student and teacher models. This strategy not only fully leverages the existing capabilities of the student model but also accelerates their mastery over task difficulties, ultimately improving the overall accuracy and robustness of the model. By building on existing theories, we introduce an innovative approach that effectively addresses the limitations of traditional curriculum learning in the context of knowledge distillation.

We define the complexity of samples based on their predictive performance across different modalities. Specifically, we classify samples into three categories based on predictions from the text modal model M_{text} and the multimodal model M_{multi}:

1. Difficult Samples: Samples that are incorrectly predicted by the single modal model but correctly predicted by the multimodal model, or vice versa:

$$Difficult_1 = \{x \mid Y_{\text{text}}(x) = False \wedge Y_{\text{multi}}(x) = True\} \tag{9}$$

$$Difficult_2 = \{x \mid Y_{\text{text}}(x) = True \wedge Y_{\text{multi}}(x) = False\} \tag{10}$$

$$Difficult = Difficult_1 \cup Difficult_2 \tag{11}$$

2. Simple Samples: Samples that are correctly predicted by both modalities:

$$Simple = \{x \mid Y_{\text{text}}(x) = True \wedge Y_{\text{multi}}(x) = True\} \tag{12}$$

3. General Samples: Samples predicted incorrectly by both models:

$$General = \{x \mid Y_{\text{text}}(x) = False, Y_{\text{multi}}(x) = False\} \tag{13}$$

In these equations, the symbols have the following meanings: - x: Represents the input sample. - $Y_{\text{text}}(x)$: Represents the prediction result of the single-modal model for sample x. - $Y_{\text{multi}}(x)$: Represents the prediction result of the multimodal model for sample x. - *True*: Indicates a correct prediction by the model. - *False*: Indicates an incorrect prediction by the model.

During training, we adopt a learning sequence that prioritizes difficult samples first, followed by moderately complex samples, and concludes with simpler samples. This progressive strategy can be represented as:

$$S = [Difficult, General, Simple] \tag{14}$$

Note that a general sample is defined as the case where both models make incorrect predictions. We place these samples in the middle of the learning sequence because their misclassification is constrained by the model architecture, making them difficult to distinguish. Since adjustments to the training order do not significantly improve prediction accuracy, our method does not focus on this aspect. This positioning neither disrupts the learning of complex samples nor compromises the performance of simpler ones.

Combined with the MIC and ILC methods, this training strategy constitutes a comprehensive framework for enhancing emotion recognition performance in multimodal speech tasks.

 doi:10.20944/preprints202601.2321.v1

3. Results

3.1. Model Training

In our proposed training framework, we define the total loss L_{total} as a combination of multiple components: the DL L_{distill}, RL L_{ref}, classification loss L_{cls}, and regularization loss L_{reg}. Specifically, L_{reg} refers to L2 regularization, which encourages the model parameters to remain small and helps prevent overfitting. This comprehensive loss function ensures balanced learning and helps optimize model performance effectively.

The total loss can be formulated as follows:

$$L_{\text{total}} = L_{\text{total\_distill}} + L_{\text{total\_ref}} + L_{\text{cls}} + L_{\text{reg}} \tag{15}$$

4. Experiment Results

4.1. Experiment Setup

For our evaluations, we utilized the MELD and IEMOCAP datasets [35,36].To ensure a comprehensive assessment of model performance, each dataset was divided into training, validation, and test sets, as shown in Table 3.

Table 3. Statistics of two datasets.

Dataset	train	val	test	Classes
MELD	9989	1109	2610	7
IEMOCAP	5810		1623	6

Our experiments were conducted on an NVIDIA RTX 3090 GPU, which provided the necessary computational power for both training and inference. The key hyperparameters of our model - α, β, γ, σ, and T -were manually tuned based on preliminary experiments to optimize performance. Their values were set to 0.5, 0.5, 0.5, 0.5, and 2, respectively.

The reported results represent the average performance over five randomly initialized runs on the test set. To facilitate reproducibility and further research, we have open-sourced the code for our experiments at https://github.com/gitxun/MRD.

4.2. Comparison with Existing Models

M3Net [18], MGLRA [19], both the baseline CMERC [12] and EmoCaps [14] are GNN methods, the MPT-HCL [15] and CFN-ESA [16] methods are transformer-based, and TelME [25] uses distillation methods. All of these methods are introduced in the Introduction section.

4.3. Result Analysis

To investigate the effectiveness of our methods, we conducted comparative experiments. In Table 4, we present comparative experiments using state-of-the-art models recently introduced in the literature, all evaluated on the IEMOCAP and MELD datasets. Our model achieves superior performance in most categories in terms of the weighted F1 score (W-F1), particularly in the overall W-F1 score, where it demonstrates a significant improvement over previous models.

Table 4. Comparison of different methods on the IEMOCAP and MELD datasets.

Methods	IEMOCAP							MELD							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
M3Net [18]	52.74	79.39	67.55	69.30	74.39	66.58	69.24	79.31	58.76	20.51	40.46	63.21	26.17	52.53	65.47
MGLRA [19]	63.50	81.50	71.50	61.10	76.30	67.80	70.10	80.80	59.50	0.00	27.80	66.50	0.00	48.40	64.90
TelME [25]	49.46	83.48	67.42	68.49	77.38	68.63	70.48	80.22	60.33	26.97	43.45	65.67	26.42	56.70	67.37
CFN-ESA [16]	53.67	80.60	71.65	70.32	74.82	68.06	71.04	80.05	58.78	21.62	41.82	66.50	26.92	54.18	66.70
EmoCaps [14]	**71.91**	85.06	64.48	68.99	78.41	66.76	71.77	77.12	**63.19**	3.03	42.52	57.50	7.69	57.54	64.00
MPT-HCL [15]	58.13	85.97	66.75	69.96	74.06	69.06	72.51	77.82	58.26	21.52	**45.15**	60.18	30.36	59.25	65.02
CMERC(baseline) [12]	60.73	81.89	71.65	69.51	77.45	67.02	71.98	80.18	60.42	24.69	40.48	65.30	32.31	54.16	66.85
MRD(ours)	68.86	**89.67**	**75.58**	**73.84**	**85.48**	**72.58**	**78.05**	**81.59**	62.72	**32.94**	43.71	**67.15**	**33.33**	**58.04**	**69.01**

 doi:10.20944/preprints202601.2321.v1

Specifically, our method outperforms the baseline model by 6.07% on the IEMOCAP dataset and 2.16% on the MELD dataset. These results highlight the effectiveness of our approach. The larger improvement observed on IEMOCAP can be attributed to the presence of more complex samples compared to the MELD dataset, which benefits more from our model's ability to handle intricate patterns in multimodal emotion recognition. Furthermore, due to the different distributions of the two datasets, the IEMOCAP category distribution is more uniform, unlike the MELD data set which has significant differences, as shown in Figures 2 and 3, resulting in a much more significant improvement in the final results achieved by IEMOCAP compared to MELD.

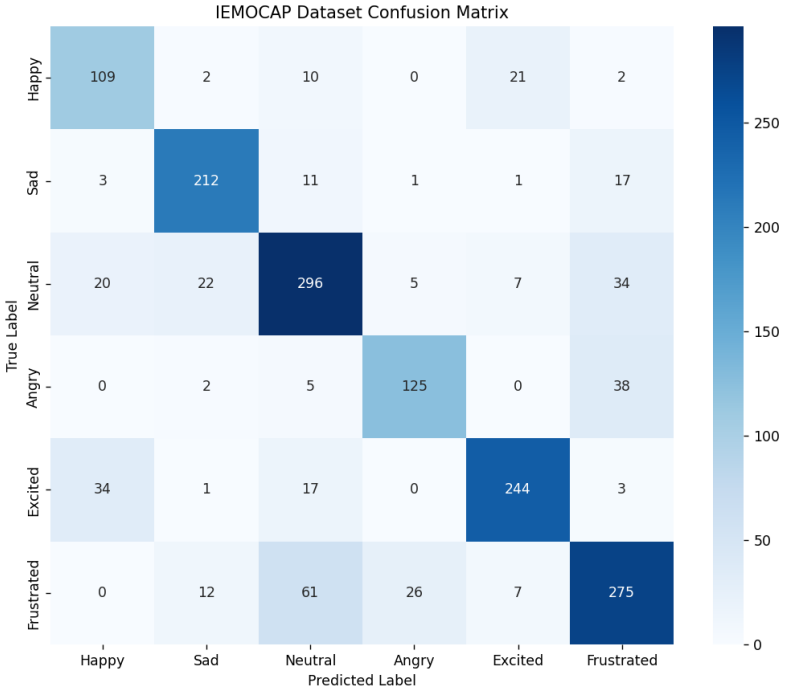

Figure 2. The Confusion Matrix of IEMOCAP Dataset

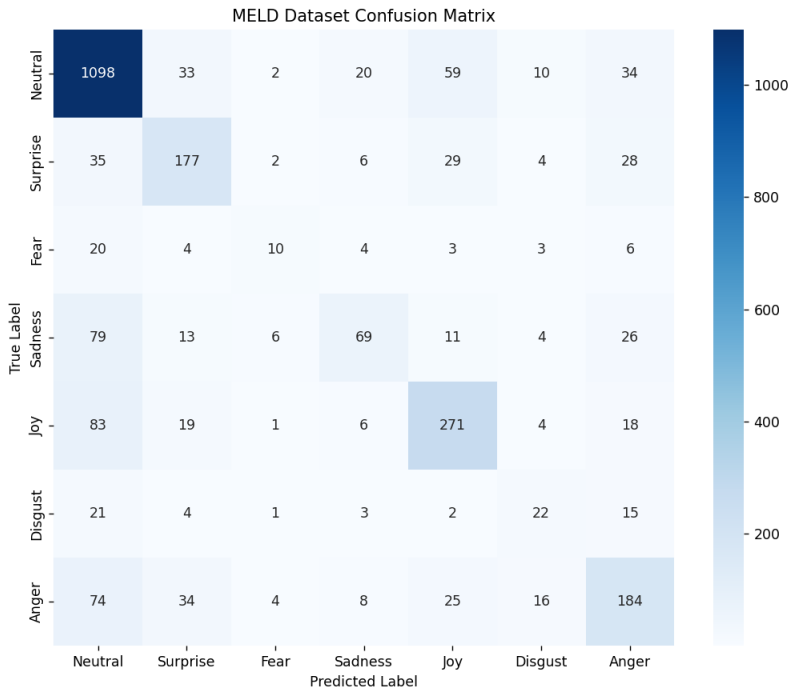

Figure 3. The Confusion Matrix of MELD Dataset

 doi:10.20944/preprints202601.2321.v1

On the IEMOCAP dataset, our model performs slightly worse than EmoCaps in classifying the Happy category. This may be due to the fact that happy emotion contains valuable information in other modalities, such as audio and visual, which our model has not fully utilized. If additional modalities were incorporated into the distillation process, the performance of this category could probably be improved. In contrast, our model outperforms the state-of-the-art methods in classifying the Sad and Excited categories. This improvement may be attributed to the high content of information such as these emotions in the text modality, which our approach effectively captures.

On the MELD dataset, our model performs slightly worse than the EmoCaps model in the Surprise category and slightly worse than the MPT-HCL model in the Sadness category. This may also be due to the lack of information from other modalities, which could be addressed by incorporating additional modalities for joint distillation. In contrast, our model achieves particularly strong performance in the Fear and Anger categories. This can be attributed to the high expressive power of the text modality for these specific emotions, enabling our model to capture their distinguishing features more effectively.

In the analysis of comparative experimental results, although MRD significantly outperforms existing methods on both the IEMOCAP and MELD datasets, the specific sources of its advantages deserve further analysis. The MRD model effectively integrates emotional information between different modalities through multimodal knowledge distillation, especially on the IEMOCAP dataset with complex samples and more balanced data distribution. This advantage is fully utilized. However, in data sets with an uneven distribution of categories such as MELD, although the overall improvement of the model is significant, there is still room for improvement in the performance of low-frequency categories (such as Fear, Dissgust, etc.), which reflects that the current distillation mechanism needs to strengthen its discriminative ability for small sample categories.

From the perspective of category performance, MRD is slightly inferior to EmoCaps in the Happy category, possibly due not only to insufficient utilization of audio and visual modality information, but also to insufficient characterization of high-frequency positive emotions in current text feature extraction methods. The confusion matrix shows that similar categories such as Happy and Excited are easily confused, indicating that the model still needs a stronger ability to capture subtle semantic and contextual changes. In the future, we can consider incorporating context-aware mechanisms and multimodal interaction modules to further enhance the model's discriminative ability. Compared to structures such as GNN and the Transformer, MRD achieves a balance between performance and efficiency through knowledge distillation, making it particularly suitable for practical applications in resource-constrained scenarios. However, currently, the main evaluation indicator is weighted F1. In order to fully reflect the adaptability of the model to various categories, it is recommended to introduce comprehensive indicators such as macro average F1 in the future.

Overall, MRD has demonstrated strong comprehensive ability and practical application potential in multimodal emotion recognition tasks, but there is still room for further improvement in addressing weak links in some categories and adaptability to small sample categories.

4.4. Ablation Study

To evaluate the effectiveness of the distilled single-text modality model, we compare its performance with the baseline model, as shown in Table 5. The distilled text modality model outperforms the baseline by 5.82% and 1.83% on the IEMOCAP and MELD datasets, respectively.

Although the single-text modality does not match the performance of the multimodal model due to the limitations of single-modality learning, it still surpasses the multimodal model in certain categories. This highlights the inherent ambiguity in classification tasks. Despite distillation, some complex samples remain challenging to classify, indicating that additional modalities could further enhance performance.

 doi:10.20944/preprints202601.2321.v1

Table 5. Performance Comparison of Ablation Studies.

Methods	IEMOCAP							MELD							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
MRD(ours)	68.86	**89.67**	75.58	73.84	**85.48**	72.58	**78.05**	**81.59**	**62.72**	**32.94**	43.71	67.15	**33.33**	**58.04**	**69.01**
MRD(text-only)	64.18	85.13	**77.31**	**76.97**	84.55	**73.78**	77.80	81.46	62.30	32.10	42.39	67.83	33.06	56.48	68.68
MRD(no-MIC)	60.73	81.89	71.65	69.51	77.45	67.02	71.98	80.18	60.42	24.69	40.48	65.30	32.31	54.16	66.85
MRD(no-ILC)	**69.08**	84.84	75.13	76.16	84.59	73.19	77.45	81.32	62.02	21.21	43.14	**68.03**	**33.85**	55.76	68.39
MRD(no-RCL)	60.73	81.17	71.78	67.63	77.95	65.51	71.45	80.18	60.42	24.69	40.48	65.30	32.31	54.16	66.85

Removing the MIC module results in no information exchange between models and no performance improvement. Likewise, eliminating the ILC module leads to an expected decline in overall performance due to overfitting and loss of model characteristics, although slight gains in specific categories may occur due to overlearning. The RCL module proves crucial for overall performance, as our hypothesis suggested. Training in simpler samples first can lead to overfitting, making it difficult for the model to learn from more complex samples later in the training process.

5. Conclusion

Our MRD method demonstrates significant effectiveness in addressing ambiguity issues in both multimodal and single-mode complex sample classification. The results of our ablation experiments further validate the contributions of our proposed three key components: MIC: Facilitates mutual learning between modalities and reduces classification ambiguity. ILC: Prevents overfitting and helps maintain the unique characteristics of each modality. RCL: Training in complex samples is prioritized, improving the model's ability to handle challenging cases.

Furthermore, a detailed analysis of each module's impact on individual emotion categories reveals additional insights. example, the removal of the MIC module particularly degrades the performance in the 'Frustrated' and 'Sad' classes on IEMOCAP, reflecting the necessity of multimodal information exchange, especially for subtle or overlapping emotions. The ILC module, while leading to slight improvements in certain frequent categories upon removal, causes overall declines and increased variance, underlining its regularization role. In particular, the RCL module is shown to be indispensable for maintaining performance in challenging, low-frequency categories, confirming its value for complex sample learning.

In addition, repeated ablation experiments with different random seeds yield consistent trends, demonstrating the robustness of these findings. These results not only validate the rationality of each component, but also provide practical guidance for model selection under varying application scenarios.

In conclusion, our ablation study demonstrates that the design of the MIC, ILC, and RCL modules is necessary and complementary, enabling the MRD framework to achieve state-of-the-art performance in multimodal emotion recognition tasks.

References

1. Schoneveld, L.; Othmani, A.; Abdelkawy, H. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters* **2021**, *146*, 1–7.
2. Ezzameli, K.; Mahersia, H. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion* **2023**, *99*, 101847.
3. Yang, Z.; Li, Z.; Zhou, S.; Zhang, L.; Serikawa, S. Speech emotion recognition based on multi-feature speed rate and LSTM. *Neurocomputing* **2024**, *601*, 128177.
4. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* **2021**, *10*, 1163.
5. Zhang, H.; Yan, Y.; Cai, Z.; Zhan, P.; Chen, B.; Jiang, B.; Xie, B. Reconstructing representations using diffusion models for multimodal sentiment analysis through reading comprehension. *Applied Soft Computing* **2024**, *167*, 112346.
6. Al-Saadawi, H.F.T.; Das, R. TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network. *Applied Sciences* **2024**, *14*, 2252.

 doi:10.20944/preprints202601.2321.v1

7. Lian, H.; Lu, C.; Li, S.; Zhao, Y.; Tang, C.; Zong, Y. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy* **2023**, *25*, 1440.
8. Khan, M.; Gueaieb, W.; El Saddik, A.; Kwon, S. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications* **2024**, *245*, 122946.
9. Le, H.D.; Lee, G.S.; Kim, S.H.; Kim, S.; Yang, H.J. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access* **2023**, *11*, 14742–14751.
10. Sun, H.; Zhao, S.; Wang, X.; Zeng, W.; Chen, Y.; Qin, Y. Fine-Grained Disentangled Representation Learning For Multimodal Emotion Recognition. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 11051–11055.
11. Li, X.; Liu, J.; Xie, Y.; Gong, P.; Zhang, X.; He, H. Magdra: a multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition. *Knowledge-Based Systems* **2024**, *283*, 111126.
12. Tu, G.; Xiong, F.; Liang, B.; Wang, H.; Zeng, X.; Xu, R. Multimodal Emotion Recognition Calibration in Conversations. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 9621–9630.
13. Sun, L.; Lian, Z.; Liu, B.; Tao, J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing* **2023**, *15*, 309–325.
14. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1610–1618.
15. Zou, S.; Huang, X.; Shen, X. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5994–6003.
16. Li, J.; Wang, X.; Liu, Y.; Zeng, Z. CFN-ESA: A Cross-Modal Fusion Network With Emotion-Shift Awareness for Dialogue Emotion Recognition. *IEEE Transactions on Affective Computing* **2024**.
17. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.V.; Modi, A. COGMEN: COntextualized GNN based multimodal emotion recognitioN. *arXiv preprint arXiv:2205.02455* **2022**.
18. Chen, F.; Shao, J.; Zhu, S.; Shen, H.T. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10761–10770.
19. Meng, T.; Zhang, F.; Shou, Y.; Shao, H.; Ai, W.; Li, K. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2024**.
20. Dai, Y.; Li, Y.; Chen, D.; Li, J.; Lu, G. Multimodal Decoupled Distillation Graph Neural Network for Emotion Recognition in Conversation. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**.
21. Li, M.; Yang, D.; Zhao, X.; Wang, S.; Wang, Y.; Yang, K.; Sun, M.; Kou, D.; Qian, Z.; Zhang, L. Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12458–12468.
22. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. Disentangled representation learning for multimodal emotion recognition. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1642–1651.
23. Yin, G.; Liu, Y.; Liu, T.; Zhang, H.; Fang, F.; Tang, C.; Jiang, L. Token-disentangling Mutual Transformer for multimodal emotion recognition. *Engineering Applications of Artificial Intelligence* **2024**, *133*, 108348.
24. Sun, T.; Wei, Y.; Ni, J.; Liu, Z.; Song, X.; Wang, Y.; Nie, L. Muti-modal Emotion Recognition via Hierarchical Knowledge Distillation. *IEEE Transactions on Multimedia* **2024**.
25. Yun, T.; Lim, H.; Lee, J.; Song, M. TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 82–95.
26. Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; Zhou, T. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116* **2024**.
27. Naik, P.; Nelaballi, S.; Pusuluri, V.S.; Kim, D.K. Deep learning-based code refactoring: A review of current knowledge. *Journal of Computer Information Systems* **2024**, *64*, 314–328.

 doi:10.20944/preprints202601.2321.v1

28. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In Proceedings of the Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.
29. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**, *364*.
30. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.
31. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
32. Hacohen, G.; Weinshall, D. On the power of curriculum learning in training deep networks. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 2535–2544.
33. Florensa, C.; Held, D.; Wulfmeier, M.; Zhang, M.; Abbeel, P. Reverse curriculum generation for reinforcement learning. In Proceedings of the Conference on robot learning. PMLR, 2017, pp. 482–495.
34. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761–769.
35. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 527–536.
36. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* **2008**, *42*, 335–359.