Coordinated Uni-modal Assistance for Enhancing Multi-modal Learning

Hongpeng Pan[1], Yang Yang[1*]

[1]Nanjing University of Science and Technology
{hongpengpancs, yyang}@njust.edu.cn

***Abstract*—Existing multi-modal learning methods often struggle with modality imbalance in real-world applications, leading to suboptimal performance. A core reason is that under the multi-modal joint learning paradigm, the dominant modality tends to control the learning process, leaving non-dominant modalities insufficiently learned and underutilized. Some attempts naturally introduce additional uni-modal objectives to alleviate this issue. However, they overlook the problem of optimization conflicts between multi-modal objectives and uni-modal objectives, which adversely impact overall performance. To address this challenge, we propose Multi-modal learning with Coordinative Uni-modal assisTance (MCUT), which incorporates collaborative uni-modal tasks alongside multi-modal tasks. Specifically, it leverages a meta-optimization approach to maximize the inner product of gradients originating from both tasks, alleviating optimization conflicts. Moreover, based on gradient fusion analysis, we enhance the assistance of uni-modal tasks by weighting the phases of meta-optimization, further boosting performance. Experiments on multiple multi-modal datasets demonstrate MCUT's superiority over existing methods. The code is available in https://github.com/njustkmg/ICME25-MCUT.**

***Index Terms*—Multi-modal Learning, Modality Imbalance, Meta-Optimization**

I. Introduction

Multi-modal information can provide a more detailed and comprehensive data representation than uni-modal sources, attracting widespread attention in multiple fields [1], [2]. The aim of multi-modal learning is to integrate information from different modalities to achieve better results than uni-modal methods, prompting the development of many multi-modal learning techniques [3], [4].

Recent studies [5], [6] have revealed a prevalent phenomenon termed "modality imbalance" in multi-modal data scenarios, where a dominant modality controls the optimization process, thereby degrading model performance. In certain cases, multi-modal learning may even perform worse than uni-modal learning [5]. To address this challenge, [6]–[8] proposed modality modulation strategies that manually or dynamically adjust modality-specific learning rates. In contrast, methods incorporating additional uni-modal objectives [9], [10] have shown superior performance. As illustrated in Figure 1, the simple introduction of additional uni-modal losses, namely Multi-Loss, can achieve performance comparable to OGM-GE, a state-of-the-art single-loss rebalancing method. According to [9], effective multi-modal learning necessitates consideration of both paired features, learned from cross-modal interactions, and uni-modal features, independently acquired from each modality. However, methods like Multi-Loss still fail to address the issue of insufficient learning by uni-modal encoders, where each modality branch performs worse than its best uni-modal result, ultimately degrading overall performance. The primary challenge arises from optimization conflicts during the joint training of multiple objectives, making it difficult to determine a set of model parameters that can effectively satisfy all objectives simultaneously.

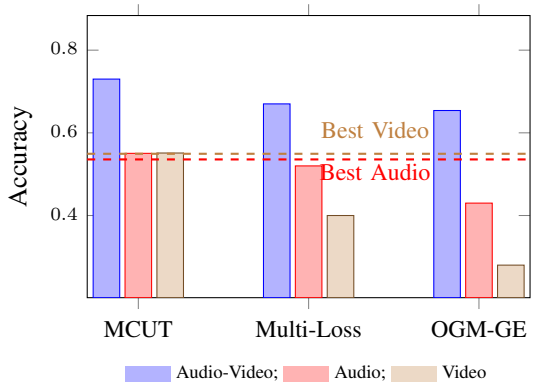

Fig. 1: Performance of the multi-modal method's predictions and its individual modality branches on the Kinetics-Sound [11] dataset, designed for action classification tasks. "Multi-Loss" is a basic approach where all losses are directly summed with equal weight. "Best Audio/Video" refers to the performance of each modality trained independently on its uni-modal encoder.

To tackle this challenge, we propose a novel framework called Multi-modal learning with Coordinative Uni-modal assisTance (MCUT), which integrates collaborative uni-modal tasks with multi-modal tasks to more effectively learn paired and uni-modal features. Specifically, we begin by treating the learning of features as two core phases within a meta-optimization framework. During meta-training phase, we focus on extracting paired features from multi-modal objectives. In meta-testing phase, we validate and optimize uni-modal features using the same samples. This approach prevents over-reliance on the modality that performs well during meta-training and ensures effective learning of uni-modal features.

*Corresponding author

Theoretically, this method mitigates interference with the learning of the original paired features by maximizing the inner product of gradients from both tasks. Moreover, based on gradient fusion analysis, we dynamically enhance the assistance of uni-modal tasks by weighting the meta-training and meta-testing phases. To accomplish this, we continuously compute performance indicators for both tasks to adjust their relative weights. Consequently, we can ensure reliable collaborative learning between paired and uni-modal features.

II. Related Work

A. Imbalanced Multi-modal Learning

Although many multi-modal methods have proven effective, [6] has identified a modality imbalance issue in multi-modal joint training, where optimization dominated by one modality leads to suboptimal model performance. Various strategies [5], [6], [9], [10], [12]–[15] have been proposed to overcome modality imbalance in multi-modal learning. [6], [12] attempt to address this by employing a dynamic gradient modulation strategy to reduce the learning rate of the dominant modality, thereby mitigating its suppression of learning from non-dominant modalities. [13] computes the cross-modal cooperation strength based on Shapley values to balance each modality. However, these methods rely on hyper-parameter tuning and primarily focus on gradient modulation under a multi-modal objective, where the learning of uni-modal features is not actively considered, resulting in suboptimal performance. [14] uses an alternating modality learning architecture, which makes it incompatible with existing fusion methods and multi-modal structures. Furthermore, [5], [9], [10] introduce additional uni-modal structures and tasks into the network to emphasize the learning of uni-modal features, but overlooks the issue of inconsistent optimization between different tasks. Therefore, in this work, we propose the MCUT algorithm, which aims to ensure paired feature learning in multi-modal tasks while collaboratively and reliably optimizing uni-modal features in uni-modal tasks.

B. Meta Learning

Meta-learning, also known as learning to learn, aims to enable models to quickly adapt to new tasks by leveraging past learning experiences. Its essence lies in the development of learning algorithms that can be widely applicable to new tasks, rather than just enhancing performance for individual tasks. MAML [16] is a milestone in this research direction, widely adopted for its ability to provide effective parameter initialization for new tasks. [17] Integrates meta-learning into domain generalization, achieving comprehensive consideration of the transition between source and target domains during training. Additionally, in cross-modal alignment scenarios, [18] applies meta-learning to optimize modal representation spaces by using strongly and weakly paired cross-modal data. [19] trains a meta-model to transfer target-specific information from the language space to the image space. In contrast to previous approaches, our work explores how to execute meta-learning frameworks for multi-modal classification tasks, enabling rapid adaptation to additional uni-modal tasks for learning uni-modal features while ensuring the effectiveness of learning original paired features. We pioneer the integration of meta-optimization with multi-modal classification tasks to tackle the challenge of modality imbalance.

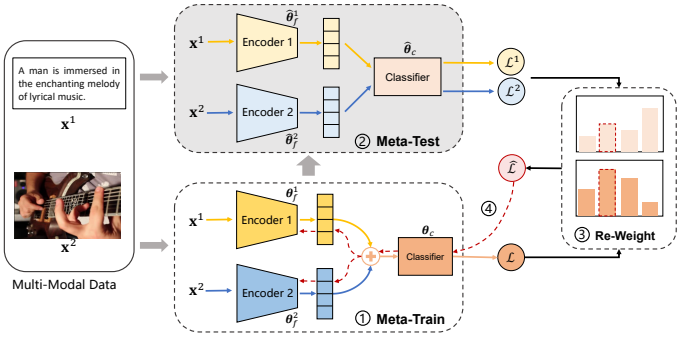

Fig. 2: Overall framework of our proposed MCUT strategy.

III. Proposed Method

A. Multi-modal Network

Let training set $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_N, \mathbf{y}_N)\}$, where each example $\mathbf{x}_i$ consists of M modalities, i.e., $\mathbf{x}_i = \{\mathbf{x}_i^m\}_{m=1}^M$, and the class label $\mathbf{y}_i \in \mathbb{R}^C$, with C representing the number of classes. The objective is to train a model using this dataset $\mathcal{D}$ to accurately predict $\mathbf{y}_i$. Most existing multi-modal joint learning methods [5], [6] typically employ a model-agnostic approach, utilizing multiple branches for final prediction. These branches consist of M feature encoders, $\{\varphi^m(\mathbf{x}_i^m; \boldsymbol{\theta}_f^m)\}_{m=1}^M$, where m-th encoder aims to extract representations from the $\mathbf{x}_i^m$ data, with learning parameters $\boldsymbol{\theta}_f^m$. Then, the multi-modal fusion operation can be denoted as $\varphi(\mathbf{x}_i) = [\varphi^1(\mathbf{x}_i^1), \varphi^2(\mathbf{x}_i^2), \cdots, \varphi^M(\mathbf{x}_i^M)]^\top$. Subsequently, the fused features are inputted into the classifier h. The output in the multi-modal model can be expressed as:

$$q(\mathbf{x}_i) = h(\varphi(\mathbf{x}_i; \boldsymbol{\theta}_f); \boldsymbol{\theta}_c), \tag{1}$$

where $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_c$ are the learning parameters representing φ and h, respectively. Finally, the objective of the vanilla multi-modal joint model is to train the $\boldsymbol{\theta}_f, \boldsymbol{\theta}_c$ to predict $\mathbf{y}$ based on $\mathbf{x}$, by minimizing the loss between the prediction and the ground truth:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^\top \log q(\mathbf{x}_i). \tag{2}$$

However, [6] discover that this simultaneous optimization of all encoders using multi-modal joint learning paradigm would be affected by modality imbalance, resulting in insufficient optimization of non-dominant modality, impairing the overall model performance. Intuitively, we can additionally incorporate uni-modal tasks to assist in learning uni-modal features. The multi-loss objectives are:

$$\underset{\boldsymbol{\theta}}{\texttt{argmin}}\ \mathcal{L}(\boldsymbol{\theta}) + \sum_{m=1}^{M} \mathcal{L}^m(\boldsymbol{\theta}^m), \tag{3}$$

where $\mathcal{L}^m(\boldsymbol{\theta}^m) = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log q^m(\mathbf{x}_i^m)$, $q^m(\mathbf{x}_i^m) = h(\varphi^m(\mathbf{x}_i^m;\boldsymbol{\theta}_f^m);\boldsymbol{\theta}_c)$ represents the prediction for the uni-modal task. For convenience, we denote the combination of parameters $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_c$ as $\boldsymbol{\theta}$, and the combination of parameters $\boldsymbol{\theta}_f^m$ and $\boldsymbol{\theta}_c$ as $\boldsymbol{\theta}^m$. Note that we use the same classifier for uni-modal predictions, and during execution, we employ zero-padding for other modality features, rendering them inactive. The objective of multi-modal joint learning is to learn paired features across multiple modalities, while uni-modal learning aims to capture uni-modal features to mitigate modality imbalance. Nevertheless, directly optimizing the average loss of these two objectives may lead to adverse effects. One reason is gradient conflict [20]. Gradients on the shared parameters may point in conflicting directions, leading to inconsistent optimization. The interference among losses reduces the ability of encoders in the multi-modal network to effectively learn paired features or uni-modal features, ultimately affecting overall performance. To address this issue, we propose MCUT, and its overall framework, as shown in Figure 2, consists of two main components: Coordinating Uni-modal Tasks and Dynamic Enhancement Assistance.

B. Coordinating Uni-modal Tasks

Drawing inspiration from meta-learning [21] strategies, we employ meta-optimization to coordinate multi-modal tasks and uni-modal tasks. Unlike previous approaches that divide samples into meta-train and meta-test sets, we execute multi-modal and uni-modal tasks separately on the same set of samples during both the meta-train (multi-modal joint learning) and meta-test (uni-modal learning) stages. Our goal is to maximize the inner product of gradients from both tasks, thereby ensuring consistent learning of paired features and uni-modal features. To elucidate the role of meta-optimization in coordinating optimization between the tasks, we will first examine the optimization equation:

$$\underset{\boldsymbol{\theta}}{\texttt{argmin}}\ \mathcal{L}(\boldsymbol{\theta}) + \beta\sum_{m=1}^{M}\mathcal{L}^m(\boldsymbol{\theta}^m - \alpha\nabla_{\boldsymbol{\theta}^m}\mathcal{L}(\boldsymbol{\theta})), \tag{4}$$

where $\nabla_{\boldsymbol{\theta}^m}\mathcal{L}(\boldsymbol{\theta})$ represents the gradient of the loss function $\mathcal{L}(\boldsymbol{\theta})$ with respect to the parameter $\boldsymbol{\theta}^m$, and α and β are hyper-parameters, used for the meta-learning rate and for weighting meta-train and meta-test, respectively. Our goal is to minimize the loss $\mathcal{L}$ obtained during meta-train and meta-test loss $\mathcal{L}^m$ obtained after one iteration of meta-train, where $\hat{\boldsymbol{\theta}}^m \leftarrow \boldsymbol{\theta}^m - \alpha\nabla_{\boldsymbol{\theta}^m}\mathcal{L}(\boldsymbol{\theta})$. Here, we refrain from utilizing higher-order gradients, as the first-order gradients suffice to effectively guide parameter updates while economizing on computational time. Intuitively, meta-test is employed to tune the parameters after meta-train, enabling the model to perform well on its tasks. In theory, we can employ a first-order Taylor expansion for the uni-modal loss during meta-test phase:

$$\mathcal{L}^m(\mathbf{z}) = \mathcal{L}^m(a) + \nabla_a\mathcal{L}^m(a)\cdot(\mathbf{z}-a), \tag{5}$$

where $\mathbf{z}$ is a vector, a is a value near $\mathbf{z}$ that serves as the base point for the expansion of the function, and $\mathcal{L}^m(a)$ is a scalar. In this context, we consider $\boldsymbol{\theta}^m - \alpha\nabla_{\boldsymbol{\theta}^m}\mathcal{L}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^m$ as $\mathbf{z}$ and a respectively. Subsequently, we obtain:

$$\mathcal{L}^m(\boldsymbol{\theta}^m - \alpha g^m) = \mathcal{L}^m(\boldsymbol{\theta}^m) + \hat{g}^m\cdot(-\alpha g^m), \tag{6}$$

where we simplify $\nabla_{\boldsymbol{\theta}^m}\mathcal{L}(\boldsymbol{\theta})$ as g^m and the gradient of $\mathcal{L}^m(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}^m$, $\nabla_{\boldsymbol{\theta}^m}\mathcal{L}^m(\boldsymbol{\theta}^m)$, as $\hat{g}^m$. The original objective function, Equation (4), can be modified to:

$$\underset{\boldsymbol{\theta}}{\texttt{argmin}}\ \mathcal{L}(\boldsymbol{\theta}) + \sum_{m=1}^{M}(\beta\mathcal{L}^m(\boldsymbol{\theta}^m) - \beta\alpha(g^m\cdot\hat{g}^m)). \tag{7}$$

This optimization objective reflects our desire to minimize both the multi-modal loss and the uni-modal loss while also aiming to maximize the inner product of gradients from both tasks. The vector dot product operation can be viewed as the modulus of two vectors multiplied by the cosine of their angle. $g^m\cdot\hat{g}^m$ suggests that if the gradient directions of the two tasks are similar, the loss will be relatively small. Hence, we can employ meta-optimization strategies to learn paired and uni-modal features in a consistent manner.

C. Gradient Fusion Analysis

In this section, we further provide an analysis of gradient fusion under gradient conflicts. Assume multi-modal objective function f_1 and uni-modal objective function f_2 share two learnable parameters $\boldsymbol{\theta}^m$ of m-th modality. Total loss degradation is calculated when updating $\boldsymbol{\theta}^m$ using the First-order Taylor expansion:

$$\Delta L = \Delta f_1^{g_1+g_2} + \Delta f_2^{g_1+g_2} \approx -\epsilon(g_1^2+g_2^2+2g_1g_2) + o(\epsilon^2), \quad g_1\cdot g_2 < 0 \tag{8}$$

where (g_1, g_2) denote $(\nabla_{\boldsymbol{\theta}^m}f_1, \nabla_{\boldsymbol{\theta}^m}f_2)$, ϵ is low learning rate. PCGrad [20], a gradient de-conflict method, enhances the performance of multi-objective models by employing gradient re-projection. Specifically, g_1 and g_2 are reformulated as $\hat{g}_1 = \left(g_1 - \frac{g_1\cdot g_2}{\|g_2\|^2}g_2\right)$ and $\hat{g}_2 = \left(g_2 - \frac{g_2\cdot g_1}{\|g_1\|^2}g_1\right)$, respectively. ΔL resulting from re-projecting two gradients is:

$$\Delta L_{g_2 \nleftarrow g_1} = -\epsilon(g_1^2+g_2^2 - \tfrac{(g_1g_2)^2}{\|g_1\|^2} - \tfrac{(g_1g_2)^2}{\|g_2\|^2} + 2g_1g_2((\cos\alpha)^2-1)) + o(\epsilon^2) \tag{9}$$

If we only reproject g_2, $\Delta L_{g_2\to g_1} = -\epsilon(g_1^2+g_2^2-\frac{(g_1\cdot g_2)^2}{\|g_1\|^2}) + o(\epsilon^2)$. The expression for $\Delta L_{g_1\to g_2}$ is formulated similarly. Given that $\|g_1\| > \|g_2\|$, $\Delta L_{g_2\nleftarrow g_1} > \Delta L_{g_2\to g_1}$ only if $\frac{\|g_2\|}{\|g_1\|} > \frac{-0.5\cos\alpha}{(\sin\alpha)^2}$; otherwise, re-projecting the smaller gradient g_2 is preferable to adjusting both gradients. Therefore, the optimal gradient fusion strategy for $\boldsymbol{\theta}^m$ under gradient conflict is:

$$g_f = \begin{cases} (1-\frac{\|g_1\|}{\|g_2\|}\cos\alpha)g_1 + (1-\frac{\|g_2\|}{\|g_1\|}\cos\alpha)g_2, & \text{if } \frac{\|g_2\|}{\|g_1\|} > \frac{-0.5(\cos\alpha)}{(\sin\alpha)^2} \\ (1-\frac{\|g_1\|}{\|g_2\|}\cos\alpha)g_1 + g_2, & \text{otherwise.} \end{cases} \tag{10}$$

The above expression indicates that a larger gradient should always be amplified, i.e., $g_f = \eta_1 g_1 + \eta_2 g_2$ with $\eta_1 > \eta_2 > 0$. Assuming better performance corresponds to relatively smaller loss and gradient magnitude, tasks with better performance should be assigned a smaller gradient weight.

D. Dynamic Enhancement Assistance

Based on the above conclusions, we re-weight each task during the training process based on its performance. To ensure a fair comparison of the model's performance on uni-modal and multi-modal tasks, we design performance evaluation indicators for multi-modal scenarios:

$$s^J = \sum_{i\in\mathcal{B}_t} \texttt{softmax}\left(q(\mathbf{x}_i)\right)_{\hat{y}_i}, \quad s^E = \sum_{i\in\mathcal{B}_t} \texttt{softmax}\left(\bar{q}(\mathbf{x}_i)\right)_{\hat{y}_i}, \tag{11}$$

where $\bar{q}(x_i)$ represents the ensemble of predictions from all uni-modal predictions, i.e., $\frac{1}{M}\sum_{m=1}^{M} q^m(\mathbf{x}^m)$, and $\hat{y}_i = \texttt{argmax}(\mathbf{y}_i)$. $\mathcal{B}_t$ represents the mini-batch of samples randomly selected at the t-th step. Subsequently, to reduce fluctuations, we employ the momentum update method, which evaluates the performance indicators by accumulating the historical contribution. We take s^J as an example. The new performance indicator can be represented as:

$$\hat{s}^J = \lambda \hat{s}^J + (1-\lambda)s^J, \tag{12}$$

where λ is the attenuation factor. Based on the performance indicators, we design the weight factor ω, which varies with training iterations, as follows:

$$\omega = \frac{2\cdot\exp\left(\hat{s}^J/\tau\right)}{\exp\left(\hat{s}^J/\tau\right)+\exp\left(\hat{s}^E/\tau\right)}, \tag{13}$$

where $\omega \in (0,2)$, τ is a scaling factor used to adjust the sensitivity of performance indicators to weight allocation. For the same $\hat{s}^J$ and $\hat{s}^E$, a higher value of τ will make ω close to 1, and vice versa. Finally, integrating with Equation (4), we replace hyper-parameter β with ω and can formulate the parameter update strategy for the MCUT model as follows:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\partial\left(\mathcal{L}(\boldsymbol{\theta}_t) + \omega\sum_{m=1}^{M}\mathcal{L}^m\left(\boldsymbol{\theta}_t^m - \alpha\nabla_{\boldsymbol{\theta}_t^m}\mathcal{L}(\boldsymbol{\theta}_t)\right)\right)}{\partial\boldsymbol{\theta}_t}. \tag{14}$$

We also provide pseudo-code in supplemental material.

IV. EXPERIMENT

In this section, we focus on the results of multi-modal classification, ablation study and optimization analysis due to page limitations. Implementation details and additional experiments, such as comparisons of multi-task strategies, intricate framework study, computational cost evaluations, and visualizations, are provided in the supplementary material.

A. Experimental Setup

Datasets and Baselines. Building on previous research addressing modality imbalance [6], [10], we validate our approach using both the Kinetics-Sound (KS) [11] and CREAM-D [22] datasets, covering audio and video modalities. To further assess our method, we expand in two directions: analyzing text-image modalities using the Multi-Modal Sarcasm-Detection (MMSD) [23] and Twitter-15 [24] datasets, and incorporating the NVGesture dataset [25] to explore fusion beyond two modalities. For NVGesture, we follow data preparation steps from [26] and conduct pre-trained training. The imbalance multi-modal learning methods include OGR-GB [5], OGM-GE [6], Greedy [27], DOMFN [28], MSLR [12], PMR [10], AGM [13],MLA [14] and Recon-Boost [29].

Evaluation Metrics. Following [6], we use accuracy (ACC) and mean Average Precision (mAP) to evaluate audio-video datasets. For the text-image and NVGesture dataset, as suggested by [23], [24], [27], we utilize accuracy (ACC) and Macro F1-score (Mac-F1). ACC measures the proportion of concordance between predicted outcomes and true labels, Mac-F1 computes the average F1 scores for each category, and mAP measures the average precision for each category.

B. Comparison with Multi-modal Methods

Comparison with imbalanced multi-modal methods. To validate the effectiveness of our approach, we compare MCUT against strategies specifically designed to overcome modality imbalance across various types of datasets. We uniformly employ prediction summation fusion in NVGesture and feature concatenation fusion in other datasets as the baseline. Considering the limitations of specific comparative methods with two modalities, we focus on the NVGesture dataset, including methods like OGR-GB, MSLR, AGM, MLA, Recon-Boost. The results are depicted in Tables I. Uni-Modal1/Uni-Modal2/Uni-Modal3 represent Audio/Video/-, Text/Img/-, and RGB/OF/Depth for different types of datasets, respectively. We have the following findings: (1) In the CREMA-D and Twitter-15 datasets, the performance of the optimal uni-modal model clearly exceeds that of the basic multi-modal joint learning (Baseline). This gap can be attributed to the widespread issue of modality imbalance. (2) All methods designed to address modality imbalance demonstrate improved performance compared to the baseline. This finding not only underscores the detrimental effects of modality imbalance on performance but also validates the effectiveness of these methods. (3) MCUT consistently achieves superior performance across nearly all metrics, highlighting the effectiveness of our approach. Notably, MCUT shows a significant improvement on the CREMA-D dataset, with an 18.54/2.42 increase in Accuracy compared to the Baseline/Second-Best approach.

Applicable to Other Fusion Strategies. We conduct a detailed investigation of both vanilla and specially-designed fusion methods, examining their performance both before and after incorporating the MCUT strategy, by observing the changes in multi-modal results as well as the outcomes of the uni-modal branches. The fusion methods include: feature concatenation (Concat), affine transformation (FiLM) [30], multi-layer LSTM (ML-LSTM) [31], channel-wise fusion (MMTM) [26], prediction summation (Sum), and prediction weighting (Weight) [3], which are further categorized into Early Fusion (Concat, FiLM, ML-LSTM), Hybrid Fusion (MMTM), and Late Fusion (Sum, Weight). As shown in Table II, the integration of MCUT into multi-modal fusion frameworks consistently enhances their performance. Furthermore, MCUT effectively reduces the performance gap across

TABLE I: Comparison between MCUT with other SOTA methods on four datasets. The best performances are highlighted in **bold**. The underscore symbol represents the second best performance.

Methods	KS		CREMA-D		MMSD		Twitter-15		NVGesture	
	ACC	mAP	ACC	mAP	ACC	Mac-F1	ACC	Mac-F1	ACC	Mac-F1
Uni-Modal1	54.12	56.69	63.17	68.61	81.36	80.65	73.67	68.49	78.22	78.33
Uni-Modal2	55.62	58.37	66.26	74.14	71.81	70.73	58.63	43.33	78.63	78.65
Uni-Modal3	-	-	-	-	-	-	-	-	81.54	81.83
Baseline	64.55	71.30	63.31	68.41	82.86	82.40	70.11	63.86	82.57	82.68
OGR-GB	67.10	71.39	64.65	68.54	83.35	82.71	74.35	68.69	82.99	83.05
OGM-GE	66.06	71.44	66.94	71.73	83.23	82.66	74.92	68.74	-	-
Greedy	66.52	72.81	66.64	72.64	-	-	-	-	-	-
DOMFN	66.25	72.44	67.34	73.72	83.56	82.62	74.45	68.57	-	-
MSLR	65.91	71.96	65.46	71.38	84.23	83.69	72.52	64.39	82.37	82.39
PMR	66.56	71.93	66.59	70.30	83.60	82.49	74.25	68.60	-	-
AGM	66.02	72.52	67.07	73.58	84.02	83.44	74.83	69.11	82.78	82.84
MLA	70.40	74.13	79.43	85.72	84.26	83.48	73.52	67.13	83.73	83.87
ReconBoost	70.85	74.24	74.84	81.24	84.37	83.17	74.42	68.34	84.13	**86.32**
MCUT	**72.67**	**78.61**	**81.85**	**88.97**	**85.14**	**84.61**	**75.12**	**69.23**	**84.43**	84.77

TABLE II: Various fusion methods combined with MCUT. † indicates that MCUT has been applied. The evaluation metric is Accuracy. The "GAP" column represents the absolute performance gap between Audio/Text and Video/Img.

Methods	CREMA-D				Twitter-15			
	Multi	Audio	Video	GAP	Multi	Text	Image	GAP
Concat	63.31	62.50	18.81	43.69	70.11	72.42	46.57	25.85
Concat†	**81.85**	**62.76**	**68.81**	**6.05**	**75.12**	**73.48**	**58.24**	**15.24**
FiLM	66.26	61.82	30.37	31.45	72.03	72.80	50.72	22.08
FiLM†	**74.05**	**63.70**	**61.15**	**2.55**	**74.34**	**72.99**	**57.85**	**15.14**
ML-LSTM	65.05	61.02	22.31	38.71	73.67	73.19	56.70	16.49
ML-LSTM †	**80.10**	**63.67**	**68.01**	**4.34**	**74.73**	**73.28**	**58.05**	**15.23**
MMTM	66.13	**60.34**	28.09	32.25		-		-
MMTM†	**76.88**	**60.34**	**64.51**	**4.17**		-		-
Sum	63.44	62.09	21.77	40.32	73.00	73.09	54.67	18.42
Sum†	**78.22**	**62.76**	**65.05**	**2.29**	**75.02**	**73.87**	**58.24**	**15.63**
Weight	66.53	62.90	26.88	36.02	72.42	72.22	56.21	16.01
Weight†	**78.09**	**63.97**	**65.45**	**1.48**	**75.89**	**73.86**	**58.63**	**15.23**

modalities and fully leverages the potential of each modality, thereby addressing the challenge of modality imbalance.

TABLE III: The component ablation experiments of the MCUT. The symbols "CUT" and "DEA" indicate whether the two components are applied during the training process. The evaluation metric is Accuracy.

Dataset	Modal	CUT ✗	DEA ✗	CUT ✓	DEA ✗	CUT ✗	DEA ✓	CUT ✓	DEA ✓
KS	Multi	66.05		71.04		68.96		**72.67**	
	Audio	52.11		54.00		53.65		**55.04**	
	Video	40.39		53.07		44.41		**55.12**	
CREMA-D	Multi	66.31		79.03		70.56		**81.85**	
	Audio	60.48		62.09		61.02		**62.76**	
	Video	48.79		65.05		50.80		**68.81**	
MMSD	Multi	83.97		84.59		84.55		**85.14**	
	Text	82.15		83.39		83.22		**83.39**	
	Img	70.25		72.06		71.27		**72.14**	
Twitter-15	Multi	73.77		74.73		74.34		**75.12**	
	Text	73.67		**74.15**		73.77		**74.15**	
	Img	53.13		57.47		56.21		**58.24**	
NVGesture	Multi	82.78		84.03		83.59		**84.43**	
	RGB	61.20		77.59		71.57		**79.87**	
	OF	63.07		78.01		72.82		**80.29**	
	Depth	71.36		**81.74**		74.27		**81.74**	

C. Ablation Study

We conduct a component ablation analysis to evaluate the significance of each components, i.e., Coordinated Uni-modal Tasks (CUT) and Dynamic Enhancement Assistance (DEA). Table III presents the experimental results. We can observe that (1) Employing CUT significantly boosts performance by effectively alleviating conflicts between paired and uni-modal feature learning through meta-optimization, thereby enhancing both multi-modal and uni-modal results. (2) The DEA module, by solely re-weighting tasks, still enhances multi-modal learning performance as it enables better gradient integration. (3) By incorporating both components, MCUT achieves the best results, demonstrating the efficacy of each component in multi-modal learning.

D. Robustness Analysis of the Pre-trained Model

TABLE IV: Test accuracy on the Sarcasm and Twitter-15 Datasets, with the encoder using the pre-trained CLIP.

Method	MMSD			Twitter-15		
	Multi	Text	Image	Multi	Text	Image
CLIP	83.11	82.15	74.82	72.52	71.75	54.48
CLIP+MLA	84.45	83.19	77.45	73.95	72.37	56.53
CLIP+MCUT	**85.14**	**83.64**	**78.12**	**74.83**	**72.90**	**63.06**

We leverage the robustness of the large-scale vision-language pre-trained model CLIP [33] on the MMSD and Twitter-15 datasets. For this purpose, the image and text encoders are replaced with the ViT-B/32 pre-trained encoders from CLIP, followed by independent fine-tuning of the model on the MMSD and Twitter-15 datasets. As shown in Table IV, "CLIP+MLA" and "CLIP+MCUT" represent the MLA-based approach and our proposed method, respectively. Based on the results, we conclude the following: (1) Both CLIP+MLA and CLIP+MCUT consistently outperform the baseline CLIP model across all metrics, effectively addressing modality imbalance; (2) Our proposed method surpasses MLA by leveraging coordinated assistance from uni-modal tasks.

E. Analysis of Optimization Process

We investigate the changes in the cosine similarity of gradients for both tasks before and after the application of

the MCUT strategy. As illustrated in Figure 3, naive learning—achieved by simply summing multiple losses—results in significant optimization inconsistencies, particularly in the early stages of training. The introduction of the MCUT strategy effectively mitigates this issue, leading to a smoother and more consistent optimization process for both tasks. This underscores the reliability and effectiveness of MCUT in enhancing multi-modal learning.

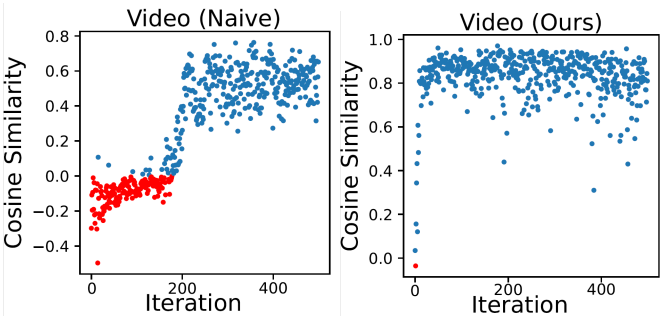

Fig. 3: The cosine similarity of gradients for the two objectives in the video encoder of CREMA-D.

V. CONCLUSION

In this study, we propose the MCUT strategy to address the issue of modality imbalance in multi-modal joint learning. Our core idea is to introduce uni-modal learning tasks that can be coordinated for optimization alongside multi-modal tasks, ensuring enhanced uni-modal feature learning while preserving paired feature learning. Specifically, we employ meta-optimization and dynamic enhancement assistance strategies to more effectively support the learning of both tasks. Finally, extensive experiments demonstrate the superiority of MCUT in alleviating modality imbalance.

VI. ACKNOWLEDGMENTS

National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), the Fundamental Research Funds for the Central Universities (No.30922010317).

REFERENCES

[1] Fengqiang Wan, Xiangyu Wu, Zhihao Guan, and Yang Yang, "Covlr: Coordinating cross-modal consistency and intra-modal relations for vision-language retrieval," in *ICME*, 2024, pp. 1–6.

[2] Zaid Khan and Yun Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *ACM MM*, 2021, pp. 3034–3042.

[3] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang, "Comprehensive semi-supervised multi-modal learning," in *IJCAI*, 2019, pp. 4092–4098.

[4] Yang Yang, Jia-Qi Yang, Ran Bao, De-Chuan Zhan, Hengshu Zhu, Xiaoru Gao, Hui Xiong, and Jian Yang, "Corporate relative valuation using heterogeneous multi-modal graph neural network," *TKDE*, vol. 35, no. 1, pp. 211–224, 2023.

[5] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in *CVPR*, 2020, pp. 12695–12705.

[6] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8238–8247.

[7] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu, "Enhancing multimodal cooperation via sample-level modality valuation," in *CVPR*, 2024, pp. 27328–27337.

[8] Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinghui Tang, "Learning to rebalance multi-modal optimization by adaptively masking subnetworks," *CoRR*, vol. abs/2404.08347, 2024.

[9] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao, "On uni-modal feature learning in supervised multi-modal learning," in *ICML*, 2023, pp. 8632–8656.

[10] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in *CVPR*, 2023, pp. 20029–20038.

[11] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *ICCV*, 2017, pp. 609–617.

[12] Yiqun Yao and Rada Mihalcea, "Modality-specific learning rates for effective multimodal additive late-fusion," in *ACL*, 2022, pp. 1824–1834.

[13] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *ICCV*, 2023, pp. 22214–22224.

[14] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao, "Multimodal representation learning by alternating unimodal adaptation," in *CVPR*, 2024, pp. 27456–27466.

[15] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu, "Facilitating multimodal classification via dynamically learning modality gap," in *NeurIPS*, 2024.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, pp. 1126–1135.

[17] Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi, "Generalizable decision boundaries: Dualistic meta-learning for open set domain generalization," in *ICCV*, 2023, pp. 11530–11539.

[18] Paul Pu Liang, Peter Wu, Ziyin Liu, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Cross-modal generalization: Learning in low resource modalities via meta-alignment," in *ACM MM*, 2021, pp. 2680–2689.

[19] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu, "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *AAAI*, 2022, pp. 1297–1305.

[20] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn, "Gradient surgery for multi-task learning," in *NeurIPS*, 2020, vol. 33, pp. 5824–5836.

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI*, 2018, pp. 3490–3497.

[22] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma, "CREMA-D: crowd-sourced emotional multimodal actors dataset," *TAC*, vol. 5, no. 4, pp. 377–390, 2014.

[23] Yitao Cai, Huiyu Cai, and Xiaojun Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *ACL*, 2019, pp. 2506–2515.

[24] Jianfei Yu and Jing Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *IJCAI*, 2019, pp. 5408–5414.

[25] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *CVPR*, 2016, pp. 4207–4215.

[26] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida, "MMTM: multimodal transfer module for CNN fusion," in *CVPR*, 2020, pp. 13289–13299.

[27] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *ICML*, 2022, pp. 24043–24055.

[28] Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu, "DOMFN: A divergence-orientated multi-modal fusion network for resume assessment," in *ACM MM*, 2022, pp. 1612–1620.

[29] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang, "Reconboost: Boosting can achieve modality reconcilement," in *ICML*, 2024.

[30] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018, pp. 3942–3951.

[31] Weizhi Nie, Yan Yan, Dan Song, and Kun Wang, "Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition," *MTA*, vol. 80, no. 11, pp. 16205–16214, 2021.