Towards Equilibrium: An Instantaneous Probe-and-Rebalance Multimodal Learning Approach

Yang Yang, **Xixian Wu**, **Qing-Yuan Jiang**[*]
Nanjing University of Science and Technology
{yyang, xixianwu, jiangqy}@njust.edu.cn

Abstract

The multimodal imbalance problem has been extensively studied to prevent the undesirable scenario where multimodal performance falls below that of unimodal models. However, existing methods typically assess the strength of modalities and perform learning simultaneously *under the imbalanced status*. This *deferred* strategy fails to rebalance multimodal learning instantaneously, leading to performance degeneration. To address this, we propose a novel multimodal learning approach, termed instantaneous probe-and-rebalance multimodal learning (IPRM), which employs a two-pass forward method to first probe (but not learn) and then perform rebalanced learning under the balanced status. Concretely, we first employ the geodesic multimodal mixup (GMM) to incorporate fusion representation and probe modality strength in the first forward phase. Then the weights are *instantaneously* recalibrated based on the probed strength, facilitating balanced training via the second forward pass. This process is applied dynamically throughout the entire training process. Extensive experiments reveal that our proposed IPRM outperforms all baselines, achieving state-of-the-art (SOTA) performance on numerous widely used datasets. The code is available at https://github.com/njustkmg/IJCAI25-IPRM.

1 Introduction

Inspired by the human ability to perceive and understand the world through multiple senses, multimodal learning (MML) [Zhao *et al.*, 2016; Yang *et al.*, 2019; Baltrusaitis *et al.*, 2019; Sun *et al.*, 2023] has emerged as a highly popular research field in recent years. In the era of deep learning, leveraging the powerful representational capabilities of deep neural networks, deep learning-based multimodal learning [Sun *et al.*, 2023] has achieved significant advancements. In real-world applications, multimodal learning has also been successfully applied across a wide range of fields [Chang *et al.*, 2015; Wang *et al.*, 2018; Zhu *et al.*, 2023; Yang *et al.*, 2024b].

Despite the abundant information provided by multimodal data, multimodal learning is expected to deliver superior performance compared to unimodal approaches. However, recent studies have revealed that multimodal models sometimes underperform compared to unimodal models in specific scenarios [Wang *et al.*, 2020]. Since different modalities vary in the amount of information they contain and their representation capabilities, their performances also tend to be inconsistent. Generally, strong modalities contain richer information and exhibit higher performance, resulting in superior outcomes, whereas weak modalities lack such advantages and perform worse [Yang *et al.*, 2015]. This phenomenon is commonly referred to as modality imbalance [Wang *et al.*, 2020; Huang *et al.*, 2022; Du *et al.*, 2022].

To deal with this problem, many efforts [Wang *et al.*, 2020; Huang *et al.*, 2022; Wu *et al.*, 2022; Zong *et al.*, 2024] have been made in recent years. Some scholars have approached this topic from a theoretical perspective, investigating issues such as generalization [Huang *et al.*, 2021], training competitiveness [Huang *et al.*, 2022], and the greedy nature [Wu *et al.*, 2022] of multimodal learning. Simultaneously, numerous multimodal learning methods have been proposed to address modality imbalance, including gradient-based approaches [Peng *et al.*, 2022; Fan *et al.*, 2023], learning rate modulation techniques [Yao and Mihalcea, 2022], and alternating optimization paradigms designed to enhance interaction [Zhang *et al.*, 2024; Fan *et al.*, 2024; Hua *et al.*, 2024]. These efforts lay a theoretical and methodological foundation for a comprehensive understanding of the nature of multimodal learning and, to some extent, mitigate performance degradation caused by modality imbalance.

Although numerous methods have been proposed to address the problem of modality imbalance, existing methods [Wang *et al.*, 2020; Li *et al.*, 2023] predominantly adopt a deferred strategy for modality rebalancing, i.e., probing the strength at first and performing multimodal learning under the imbalanced status. Among existing methods, some of them, e.g., OGR-GB [Wang *et al.*, 2020] and OGM [Peng *et al.*, 2022], first explicitly calculate multimodal imbalance-related metrics like OGM score, and then make adjustments based on metrics. While others [Li *et al.*, 2023; Zhang *et al.*, 2024; Zong *et al.*, 2024] leverage key quantities like gradients or

[*] Corresponding author.

logits that implicitly reflect the learning state of one modality to guide interventions in another. Regardless of the method, the rebalancing learning process is based on *the current imbalanced status*, rather than the balanced status after instantaneous adjustment. This deferred rebalancing strategy has an inherent limitation: it addresses modal imbalance only after it has occurred, mitigating its impact but failing to prevent the problem from arising in the first place.

Can we design an instantaneous modality rebalancing method? The answer is yes. We propose a simple two-pass forward strategy: the first forward pass probes the modality strength and recalibrates the status weights *instantaneously*, while the second forward pass applies these calibrated weights to perform model learning under the balanced status. Specifically, we first utilize geodesic multimodal mixup [Oh *et al.*, 2023], a powerful tool to capture heterogeneous representation within hypersphere, to establish the relationship across modalities with fusion representations. With this architecture, we can effortlessly adjust the modality strength between different modalities. Subsequently, we perform the first forward pass to probe the strength of each modality based on information entropy. Using this strength, we adjust the intervention intensity for different modalities in the mixup process. With this adjustment, the fusion of different modalities achieves a more balanced status. After that, the second forward pass is conducted to finalize the forward computation, loss calculation, gradient backpropagation, and parameter updates, all under a relatively balanced status. This two-pass forward process, involving evaluation followed by balanced learning, is consistently applied throughout the entire training procedure. To this end, this approach enables instantaneous multimodal rebalancing, maximizing the potential for improving the model's overall performance. Our contributions are listed as follows:

- A novel fusion representation strategy using geodesic multimodal mixup is proposed, enabling the dynamic adjustment of fusion weights based on modality strength.
- A novel two-pass forward strategy is proposed, which first probes the strength of modality imbalance and then facilitates multimodal learning under the balanced status in the second forward pass.
- A novel learning method that incorporates a two-pass forward approach into the entire training process is proposed, dynamically performing both the evaluation and balanced multimodal learning stages.
- Experiments reveal that our IPRM can achieve the best performance on various datasets by comparing with SOTA baselines.

2 Related Work

2.1 Rebalanced Multimodal Learning

Recently, multimodal learning has been observed to be less effective than unimodal models for certain tasks such as multimodal classification [Wang *et al.*, 2020; Peng *et al.*, 2022; Zong *et al.*, 2024]. To mitigate this gap and fully leverage complementary representations of different modalities, various approaches [Wang *et al.*, 2020; Peng *et al.*, 2022; Li *et al.*, 2023; Hua *et al.*, 2024; Zhang *et al.*, 2024] have been proposed.

Several approaches attempt to address this problem from a theoretical standpoint. [Huang *et al.*, 2021] demonstrates that leveraging multiple modalities leads to a lower population risk compared to utilizing only a subset of those modalities. [Huang *et al.*, 2022] examines this issue through the lens of joint training. In multimodal learning, models corresponding to different modalities tend to compete during training, causing the encoder network to focus on learning only a subset of the modalities. [Wu *et al.*, 2022] investigates the inherent greediness of deep models in multimodal learning scenarios and introduces an algorithm designed to balance the conditional learning rates across modalities during training.

Other attempts [Peng *et al.*, 2022; Li *et al.*, 2023; Hua *et al.*, 2024; Zhang *et al.*, 2024] focus on designing specific algorithms to address this issue. [Peng *et al.*, 2022; Fan *et al.*, 2023; Li *et al.*, 2023] propose algorithms based on gradient modulation to balance the learning performance across different modalities. [Yao and Mihalcea, 2022] seeks to achieve consistency in learning speeds by dynamically adjusting the learning rates of models across different modalities. [Yang *et al.*, 2025] defines an adaptive element-wise score function for parameter updates based on modal significance to rebalance optimization across modalities. [Zhang *et al.*, 2024; Hua *et al.*, 2024; Fan *et al.*, 2024] design an alternating learning paradigm to enhance the interaction across different modalities during training.

2.2 Mixup

Mixup is a widely used data augmentation strategy proposed by [Zhang *et al.*, 2018], aiming at improving model generalization and robustness. It generates virtual training samples by linearly interpolating between pairs of examples and their corresponding labels. Some variants of mixup have been proposed to improve the original mixup strategy. For instance, CutMix [Yun *et al.*, 2019] proposes to cut and paste the patches into images to improve the robustness of model against input corruptions. Manifold mixup [Verma *et al.*, 2019] design to improve the generalization of deep neural networks by applying linear interpolation not in the input space, but in the hidden representations of the model. Additionally, several mixup strategies [Fang *et al.*, 2022; Oh *et al.*, 2023] for multimodal data have also been proposed. Representative geodesic multimodal mixup [Oh *et al.*, 2023] provides a powerful tool to establish the correlation relationship across modalities, capturing the heterogeneous representation within hypersphere.

3 Multimodal Learning

Without loss of generality, we use audio and video modalities as illustrative examples. Notably, IPRM can be readily extended to scenarios involving more than two modalities. We suppose that the multimodal data training set with K category labels is defined as: $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_t}$, where $\boldsymbol{x} = (\boldsymbol{x}^a, \boldsymbol{x}^v)$ respectively denote the audio and video data points, N_t is the number of training data, and $\boldsymbol{y}_i \in \{1, \cdots, K\}$ denotes the i-th category label.

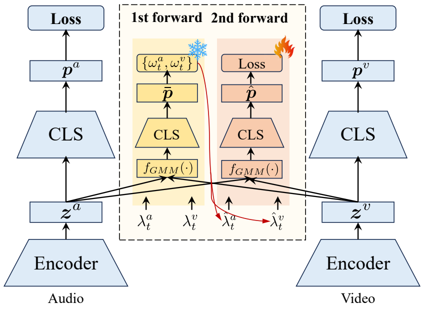


Figure 1: The architecture of our method.

Multimodal learning aims to train a model that maps data points into label space, and approximates its label as accurately as possible. In deep learning-based multimodal learning, deep neural networks serve as the foundational components of the model. Concretely, we employ g_a and g_v to denote the encoders for the audio and video modalities. Given $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_i^v$, the representations can be calculated by:

$$\forall o \in \{a, v\},\ \boldsymbol{z}_i^o = g_o(\boldsymbol{x}_i^o; \Theta_o),$$

where Θ_o denotes the parameters. Then we take a fusion function to obtain fusion representation:

$$\boldsymbol{z}_i = f(\boldsymbol{z}_i^a, \boldsymbol{z}_i^v),$$

where f denotes the fusion function like sum or averaging.

Following [Fan *et al.*, 2023] and [Wei and Hu, 2024], we consider to minimize both unimodal and multimodal losses. Hence, we define the classifiers for unimodal and multimodal branches as h_a, h_v and h, respectively. The predictions can be calculated by:

$$\forall o \in \{a, v\}, \boldsymbol{p}_i^o = \mathit{softmax}(h_o(\boldsymbol{z}_i^o; \Phi_o)),$$
$$\boldsymbol{p}_i = \mathit{softmax}(h(\boldsymbol{z}_i; \Phi)),$$

where Φ_o and Φ denote the parameters of the classifier.

Then the unimodal and multimodal loss functions can be formed as:

$$\forall o \in \{a, v\},\ \ell_u(\boldsymbol{x}_i^o, \boldsymbol{y}_i) = -\sum\nolimits_{k=1}^{K} \mathbb{1}_{k=\boldsymbol{y}_i} \log(\boldsymbol{p}_{ik}^o),$$
$$\ell_m(\boldsymbol{x}_i, \boldsymbol{y}_i) = -\sum\nolimits_{k=1}^{K} \mathbb{1}_{k=\boldsymbol{y}_i} \log(\boldsymbol{p}_{ik}),$$

where $\boldsymbol{p}_{ik}$ denotes the k-th element of $\boldsymbol{p}_i$, and $\mathbb{1}_{condition}$ denotes the indicator function, i.e., $\mathbb{1}_{true} = 1$ and $\mathbb{1}_{false} = 0$.

The overall loss can be formed as:

$$\ell(\boldsymbol{x}_i, \boldsymbol{y}_i) = \ell_m(\boldsymbol{x}_i, \boldsymbol{y}_i; \Phi) + \sum\nolimits_{o \in \{a,v\}} \ell_u(\boldsymbol{x}_i^o, \boldsymbol{y}_i; \Theta_o, \Phi_o).$$

Then, the gradient is computed through the backward process, and the parameters are updated accordingly.

4 Methodology

The architecture of IPRM is presented in Figure 1. We first introduce geodesic multimodal mixup to bridge the fusion representation. Then, we design a two-pass forward strategy to probe the strength of modality imbalance first and then perform rebalanced learning. Finally, a dynamic learning algorithm is proposed to impose the two-pass forward strategy into the entire training process.

4.1 Multimodal Fusion with GMM

Instead of linear mixup, we utilize geodesic multimodal mixup strategy to construct heterogeneous representations across modalities. Specifically, we first generate normalized representation for both audio and video modalities:

$$\forall o \in \{a, v\},\ \bar{\boldsymbol{z}}_i^o = \frac{\boldsymbol{z}_i^o}{\|\boldsymbol{z}_i^o\|_2}.$$

Then the geodesic multimodal mixup strategy is imposed on these vectors, and we can redefine the fusion function as:

$$f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \lambda) = \frac{\mathit{sin}(\lambda\theta)}{\mathit{sin}(\theta)} \bar{\boldsymbol{z}}_i^a + \frac{\mathit{sin}((1-\lambda)\theta)}{\mathit{sin}(\theta)} \bar{\boldsymbol{z}}_i^v, \quad (1)$$

where $\theta = \mathit{arccos}(\langle \bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v \rangle)$ and λ is a parameter. Symbol $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors.

After obtaining the fusion representation with geodesic multimodal mixup, we can further feed them into classification networks and calculate the prediction. Based on prediction, the multimodal loss is calculated correspondingly.

We use λ to control the influence strength of different modalities, thereby achieving the goal of characterizing intervention based on the strength of each modality.

4.2 Instantaneous Probe-and-Rebalance for MML

We then present the two-pass forward strategy in detail. This strategy can be divided into instantaneous probing phase and rebalanced learning phase.

Instantaneous Probing Phase: At the t-th iteration, given a batch of data points, we first utilize encoders to extract features. Then, during the instantaneous probing phase, we can probe the strength of modality imbalance based on multimodal and unimodal predictions after we obtain the fusion representation $\bar{\boldsymbol{z}}_i$ and prediction $\bar{\boldsymbol{p}}_i$:

$$\bar{\boldsymbol{z}}_i = f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \lambda_t^a),\ \ \bar{\boldsymbol{p}}_i = \mathit{softmax}(h(\bar{\boldsymbol{z}}_i)),$$

where λ_t^a denotes the weight controlling the strength of the modality before learning for audio, and $\lambda_t^v = 1 - \lambda_t^a$.

Then, we utilize a widely used measure Kullback–Leibler (KL) divergence to evaluate the strength of each modality:

$$\forall o \in \{a, v\}, \mathcal{D}_{\mathrm{KL}}(\mathcal{P}^o | \bar{\mathcal{P}};\ \mathcal{T}_t) = \sum\nolimits_{\boldsymbol{x}_i \in \mathcal{T}_t} \boldsymbol{p}_i^o \log\left(\frac{\boldsymbol{p}_i^o}{\bar{\boldsymbol{p}}_i}\right),$$

where $\mathcal{P}^o$ and $\bar{\mathcal{P}}$ denote the corresponding prediction distributions. KL divergence quantifies the distance between different distributions. A smaller value indicates that the two distributions are closer. Therefore, a modality with a large KL divergence is considered a weak modality, and its weight should be reduced during the hybrid enhancement process. Conversely, a modality with a small KL divergence is a strong modality, and its weight should be increased. Hence, we define the instantaneous strength weight of a specific modality based on the proportion of the KL divergence from another modality:

$$\omega_t^a \triangleq \frac{\mathcal{D}_{\mathrm{KL}}(\mathcal{P}^v | \mathcal{P};\ \mathcal{T}_t)}{\mathcal{D}_{\mathrm{KL}}(\mathcal{P}^a | \mathcal{P};\ \mathcal{T}_t) + \mathcal{D}_{\mathrm{KL}}(\mathcal{P}^v | \mathcal{P};\ \mathcal{T}_t)},$$
$$\omega_t^v \triangleq 1 - \omega_t^a. \quad (2)$$

Algorithm 1: The IPRM learning algorithm.

```
Input  : Training set $\mathcal{T}$.
Output: Learned parameters of all models.
INIT Initialize parameters $\{\Theta_a, \Theta_v, \Phi, \Phi_a, \Phi_v\}$, maximum
  iterations $M_t$, learning rate $\eta_\alpha$, modality weight
  $\omega_0^a = \omega_0^v = 0.5$, $\lambda_1^a$ and $\lambda_1^v$ based on Eq. (4).
for $t = 1$ to $M_t$ do
    Sample a mini-batch data samples $\mathcal{T}_t$.
    Calculate features based on $g_a(\cdot)$ and $g_v(\cdot)$.
    /* The first forward phase.  */
    Calculate fused feature $\bar{\boldsymbol{z}}$ the by the first forward phase.
    Calculate the instantaneous strength score based on
      Eq. (2).
    Calculate the balanced weight based on Eq. (3).
    /* The second forward phase.  */
    Calculate the prediction $\bar{\boldsymbol{p}}$ by the second forward phase.
    Calculate the gradients based on backward phase.
    Update the network parameters based on SGD.
    Update $\lambda_t^a$ and $\lambda_t^v$ based on Eq. (4).
end
```

Rebalanced Learning Phase: As the instantaneous strength score accurately captures the imbalance degree among different modalities, we directly use this parameter to update the balanced weights for each modality at t-th iteration:

$$\forall o \in \{a, v\}, \hat{\lambda}_t^o = \omega_t^o. \tag{3}$$

Then, $\hat{\lambda}_t^a$, $\hat{\lambda}_t^v$ are used to perform the second forward phase to obtain fusion representation *under the balanced status*:

$$\hat{\boldsymbol{z}}_i = f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \hat{\lambda}_t^a), \;\; \hat{\boldsymbol{p}}_i = softmax(h(\hat{\boldsymbol{z}}_i)).$$

The multimodal loss can be calculated by:

$$\hat{\ell}_m(\boldsymbol{x}_i, \boldsymbol{y}_i) = -\sum\nolimits_{k=1}^{K} \mathbb{1}_{k=\boldsymbol{y}_i} \log(\hat{\boldsymbol{p}}_{ik}).$$

Accordingly, we derive the gradients of the parameters and update them to finish the model training.

After completing the learning in t-th iteration, we update the initial weights for the next iteration to adjust the intervention intensity between modalities accordingly. We utilize the exponential moving average (EMA) to update the weight:

$$\forall o \in \{a, v\}, \lambda_{t+1}^o = \begin{cases} \omega_t^o, & t = 0, \\ \alpha \lambda_t^o + (1-\alpha)\omega_t^o, & t > 0. \end{cases} \tag{4}$$

Here, α is a hyper-parameter to tune the weight, and ω_0^o is initially set to be 0.5 in practice.

Learning Algorithm: Finally, we briefly summarize our algorithm. We train the model using the entire training dataset $\mathcal{T}$ in a mini-batch style. At t-th iteration, we first utilize encoders to extract features. And then we perform the first forward phase to obtain the instantaneous strength weight ω_t^a and ω_t^v. Based on instantaneous strength weight ω_t^a and ω_t^v, we can update the modality weight λ_t^a and λ_t^v as $\hat{\lambda}_t^a$ and $\hat{\lambda}_t^v$. In the second forward phase, the models are updated under balanced status by using $\hat{\lambda}_t^a$ and $\hat{\lambda}_t^v$. The whole algorithm is summarized in Algorithm (1).

5 Experiments

5.1 Experimental Setup

Datasets: We utilize five datasets for experiments, i.e., *CREMA-D* [Cao *et al.*, 2014], *KSounds* [Arandjelovic and Zisserman, 2017], *NVGesture* [Molchanov *et al.*, 2016], *IEMOCAP* [Busso *et al.*, 2008], and *Sarcasm* [Cai *et al.*, 2019] datasets. Among these datasets, the *CREMA-D* and *KSounds* datasets involve audio and video modalities. *NVGesture* consists of RGB, optical flow (OF), and Depth modalities. *IEMOCAP* is also a trimodal dataset, which contains audio, video and text modalities. *Sarcasm* dataset is an image-text dataset.

The *CREMA-D* dataset contains 7,442 clips from 91 actors. And it is divided into a training set with 6,698 samples and a testing set with 744 samples. The *KSounds* dataset is divided into a training set with 15K samples, a validation set with 1.9K samples, and a testing set with 1.9K samples. For *NVGesture* dataset, it is split as 1,050 data points for training and 482 for testing. And *IEMOCAP* dataset is split as a training set with 3,318 samples and a testing set with 1,107 samples. *Sarcasm* dataset consists of 24,635 and is split as a training set with 19,816 samples, a testing set with 2,409, and a validation set with 2,410 samples.

Baselines: Considering that geodesic multimodal mixup can be treated as a modified fusion function, we first utilize unimodal approaches and naive fusion strategies including Concat, Sum and Weight for comparison. Furthermore, a wide range of SOTA rebalanced multimodal learning approaches are used for comparison. They are OGR-GB [Wang *et al.*, 2020], MSLR [Yao and Mihalcea, 2022], OGM [Peng *et al.*, 2022], PMR [Fan *et al.*, 2023], AGM [Li *et al.*, 2023], MMPareto [Wei and Hu, 2024], Reconboost [Hua *et al.*, 2024], MLA [Zhang *et al.*, 2024], and LFM [Yang *et al.*, 2024a].

Evaluation Metrics: Following [Zhang *et al.*, 2024; Yang *et al.*, 2024a], the accuracy and mean average precision (MAP) are used to evaluate the performance on *CREMA-D* and *KSounds* datasets. Furthermore, for the remaining datasets, we utilize accuracy and macro F1 (Macro-F1) for evaluation. The accuracy evaluates the degree of agreement between predictions and ground-truth labels. The MAP is obtained by averaging the average precision across all categories, while the macro-F1 is determined by computing the mean of the F1 scores for each category.

Implementation Details: Following [Peng *et al.*, 2022], we adopt ResNet18 [He *et al.*, 2016] as backbones for both audio and video modalities on *CREMA-D* and *KSounds* datasets. And the models are trained from scratch. For *NVGesture* dataset, we use I3D [Carreira and Zisserman, 2017] as unimodal backbone, as described in [Wu *et al.*, 2022]. Following [Zhang *et al.*, 2024], we respectively employ large pretrained model M3AE [Geng *et al.*, 2022] and CAV-MAE [Gong *et al.*, 2023] as encoders for image/text and audio modalities on *IEMOCAP* dataset. For *Sarcasm*, ResNet50 is utilized for the image modality and BERT [Devlin *et al.*, 2019] for the text modality, which is consistent with [Yang *et al.*, 2024a]. The optimization algorithm for the audio-video and trimodal datasets is stochastic gradient descent (SGD), while Adam is employed for the image-text dataset. The

Dataset	Metric	Unimodal			Naive Fusion			IPRM
		A/A/R/A/I	V/V/O/V/T	D/T	Concat	Sum	Weight	
CREMA-D	Accuracy	63.17%	45.83%	N/A	63.61%	63.44%	66.53%	**84.27%** (↑17.74%)
	MAP	68.61%	58.79%	N/A	68.41%↓	69.08%	71.34%	**90.66%** (↑19.32%)
KSounds	Accuracy	54.12%	55.62%	N/A	64.55%	64.90%	65.33%	**74.37%** (↑9.04%)
	MAP	56.69%	58.37%	N/A	71.30%	71.03%	71.10%	**80.63%** (↑9.33%)
NVGesture	Accuracy	78.22%	78.63%	81.54%	82.37%	80.50%↓	78.42%↓	**85.89%** (↑3.52%)
	Macro-F1	78.33%	78.65%	81.83%	82.70%	80.67%↓	79.39%↓	**86.34%** (↑3.64%)
IEMOCAP	Accuracy	58.45%	30.71%	70.55%	75.97%	76.06%	69.29%↓	**80.22%** (↑4.16%)
	Macro-F1	58.29%	11.75%	69.93%	75.88%	76.03%	68.91%↓	**80.63%** (↑4.60%)
Sarcasm	Accuracy	71.81%	81.36%	N/A	82.86%	82.94%	82.65%	**85.14%** (↑2.20%)
	Macro-F1	70.73%	80.56%	N/A	82.40%	82.47%	82.19%	**84.41%** (↑1.94%)

Table 1: Performance comparison with vanilla MML. The best and the second best results are denoted as bold and underline, respectively. The symbol ↓ indicates the MML result which underperforms the best unimodal result.

Dataset	Metric	OGR-GB	MSLR	OGM	PMR	AGM	MMPareto	ReconBoost	MLA	LFM	IPRM
CREMA-D	Accuracy	64.65%	68.68%	66.12%	66.59%	67.33%	74.87%	75.57%	79.43%	83.62%	**84.27%** (↑0.65%)
	MAP	73.92%	74.12%	73.72%	70.58%	78.07%	85.35%	81.40%	85.72%	90.06%	**90.66%** (↑0.60%)
KSounds	Accuracy	67.22%	67.56%	65.82%	66.75%	67.91%	70.00%	68.55%	70.04%	72.53%	**74.37%** (↑1.84%)
	MAP	72.74%	72.82%	71.59%	72.74%	73.88%	78.50%	76.62%	79.45%	78.97%	**80.63%** (↑1.66%)
NVGesture	Accuracy	82.99%	82.37%	N/A	N/A	82.79%	83.82%	83.86%	83.40%	84.36%	**85.89%** (↑1.53%)
	Macro-F1	83.05%	82.84%	N/A	N/A	82.84%	84.24%	84.34%	83.72%	84.68%	**86.34%** (↑1.66%)
IEMOCAP	Accuracy	70.10%	76.69%	N/A	N/A	77.51%	77.69%	76.87%	79.31%	78.41%	**80.22%** (↑0.91%)
	Macro-F1	69.90%	76.77%	N/A	N/A	77.29%	77.89%	77.08%	79.73%	78.51%	**80.63%** (↑0.90%)
Sarcasm	Accuracy	82.86%	84.39%	83.60%	83.10%	83.06%	83.48%	84.37%	84.26%	84.97%	**85.14%** (↑0.17%)
	Macro-F1	82.15%	83.78%	82.93%	82.56%	82.93%	82.84%	83.17%	83.48%	**84.57%**	84.41% (↓0.16%)

Table 2: Performance comparison with SOTA rebalanced multimodal learning approaches. The best and the second best results are denoted as bold and underline, respectively.

learning rate is set to 10^{-2} for the audio-video datasets and *NVGesture*, 10^{-3} for *IEMOCAP*, and 10^{-4} for *Sarcasm*, respectively. It is then reduced by a factor of 10 when the loss saturates. The batch size is set to be 64 for *CREMA-D*, *KSounds* and *Sarcasm*, while is respectively set to be 2 and 16 for *NVGesture* and *IEMOCAP* due to out-of-memory issue. Furthermore, the hyper-parameter α is set to 0.8 for audio-video datasets and 0.7 for trimodal and image-text datasets, based on cross-validation strategy. Unless otherwise specified, the mixup in our experiments all uses the paired sample method, and we will discuss the impact of this strategy in subsection 5.5. All experiments are conducted on GeForce RTX 4090 NVIDIA card.

5.2 Main Results

We first compare our method with unimodal approaches and naive multimodal learning with different fusion functions, i.e., concat, sum, and weight [Yang *et al.*, 2019]. The results are reported in Table 1, where "A/A/R/A/I" and "V/V/O/V/T" denote the abbreviations of the modalities used for evaluation across all datasets, and "D/T" respectively denote the abbreviations of depth (for *NVGesture* dataset) and text (for *IEMOCAP* dataset). The results in Table 1 demonstrate that: (1). The multimodal learning approach outperforms the single-modal approach in most cases but performs worse in certain scenarios, which is indicated by the symbol "↓"; (2). Our IPRM can achieve the best results in all cases by a large margin, demonstrating that it is not merely a simple fusion variant but an algorithm specifically designed for the practical requirements of multimodal learning.

Dataset	w/ L-Mixup	w/o EMA	One-Pass	IPRM
CREMA-D	75.53%	83.06%	83.47%	**84.27%**
KSounds	71.94%	73.91%	73.64%	**74.37%**
NVGesture	84.85%	85.27%	84.44%	**85.89%**
IEMOCAP	75.79%	78.05%	77.60%	**80.22%**
Sarcasm	84.52%	84.81%	84.10%	**85.14%**

Table 3: Ablation study on all datasets. The best results are shown in bold.

Then we compare our IPRM with various existing SOTA rebalanced multimodal learning approaches. Table 2 provides the corresponding results, where the results of OGM and PMR on *NVGesture* and *IEMOCAP* datasets are denoted as "N/A" as these approaches cannot be applied to the case with three modalities. The results in Table 2 demonstrate that our IPRM surpasses the existing state-of-the-art algorithm and achieves superior performance in almost all cases, demonstrating that the module we designed effectively addresses the challenge of modality imbalance.

5.3 Ablation Study

We study the impact of key components of our IPRM, including geodesic multimodal mixup, EMA strategy, and two-pass forward phase. Specifically, we report the accuracy to compare the IPRM with its variants. The variants include the vari-

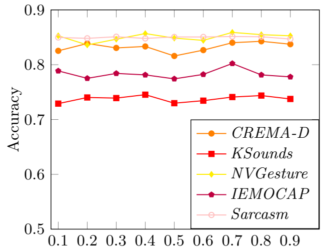


Figure 2: Sensitivity to α.

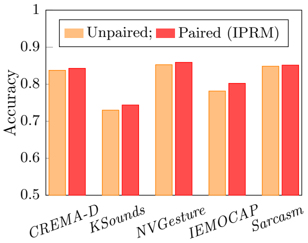


Figure 3: Unpaired GMM.

ant using linear mixup, the variant without EMA strategy, and the variant with one-pass forward phase, which are denoted as “w/ L-Mixup”, “w/o EMA”, and “One-Pass”, respectively. For One-Pass variant, the loss calculation, gradient calculation by backward, and parameter update are performed in the first forward phase, and the ω_t^a and ω_t^v are used to update λ_{t+1}^a and λ_{t+1}^v at the next iteration.

The accuracy on all datasets is reported in Table 3. The results reveal that both geodesic multimodal mixup, EMA strategy and two-pass forward phase can boost performance. On *CREMA-D*, *KSounds*, and *IEMOCAP* datasets, the most significant performance improvement is attributed to the use of geodesic multimodal mixup. In contrast, for the remaining datasets, the module contributing the greatest impact on performance is two-pass forward phase.

5.4 Sensitivity to Hyper-Parameter

We further explore the sensitivity to the hyper-parameter α. Parameter $\alpha \in (0, 1)$ is the weight of the EMA strategy in Equation (4). The accuracy on all datasets is reported in Figure 2, where α is chosen from the list $[0.1, 0.2, \cdots, 0.9]$. The accuracy in Figure 2 demonstrates that IPRM is not sensitive to the hyper-parameter α in a large range.

5.5 Further Analysis

GMM with Unpaired Sampling: In IPRM, we utilize paired multimodal data to perform model training. Thus we do not discuss the fusion strategy for category label when we apply the geodesic multimodal mixup [Oh *et al.*, 2023] in our method. In practical scenarios, unpaired multimodal data may often be used for training. Therefore, this section discusses the impact of paired and unpaired sampling strategies on the overall performance. For geodesic multimodal mixup, given an unpaired multimodal data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), (\boldsymbol{x}_j, \boldsymbol{y}_j)\}$, the augmented ground-truth can be formulated by linear combination, i.e., $\boldsymbol{y}^{'} = \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_j$. We exploit the influence of geodesic multimodal mixup with unpaired sampling. The accuracy on all datasets is presented in Figure 3. From Figure 3, we can see that across all datasets, the overall performance of the unpaired sampling method is slightly inferior to that of the paired sampling method. This performance gap can be attributed to three factors: (1). Unpaired sampling disrupts the exploration of complementary multimodal information; (2). Mixing labels introduces additional noise into the training process; (3) Unlike standard mixup, IPRM’s λ is not randomly generated from a distribution, which may shift the distribution of the generated labels towards the stronger modality.

Method	Accuracy	Training time (second/epoch)
Naive MML	63.61%	55.08 ± 0.2729
MLA	79.43%	71.12 ± 0.7025
LFM	83.62%	60.14 ± 0.0920
IPRM	84.27%	57.03 ± 0.2138

Table 4: Computation cost comparison on *CREMA-D* dataset.

Dataset	Modality	Single-CLS	Tri-CLS
NVGesture	RGB	**78.84%**	77.80%
	OF	79.25%	**81.12%**
	Depth	**82.78%**	82.16%
	Multi	**85.89%**	**85.89%**
IEMOCAP	Audio	**58.27%**	54.20%
	Video	**32.07%**	30.80%
	Text	**71.91%**	**71.91%**
	Multi	78.95%	**80.22%**

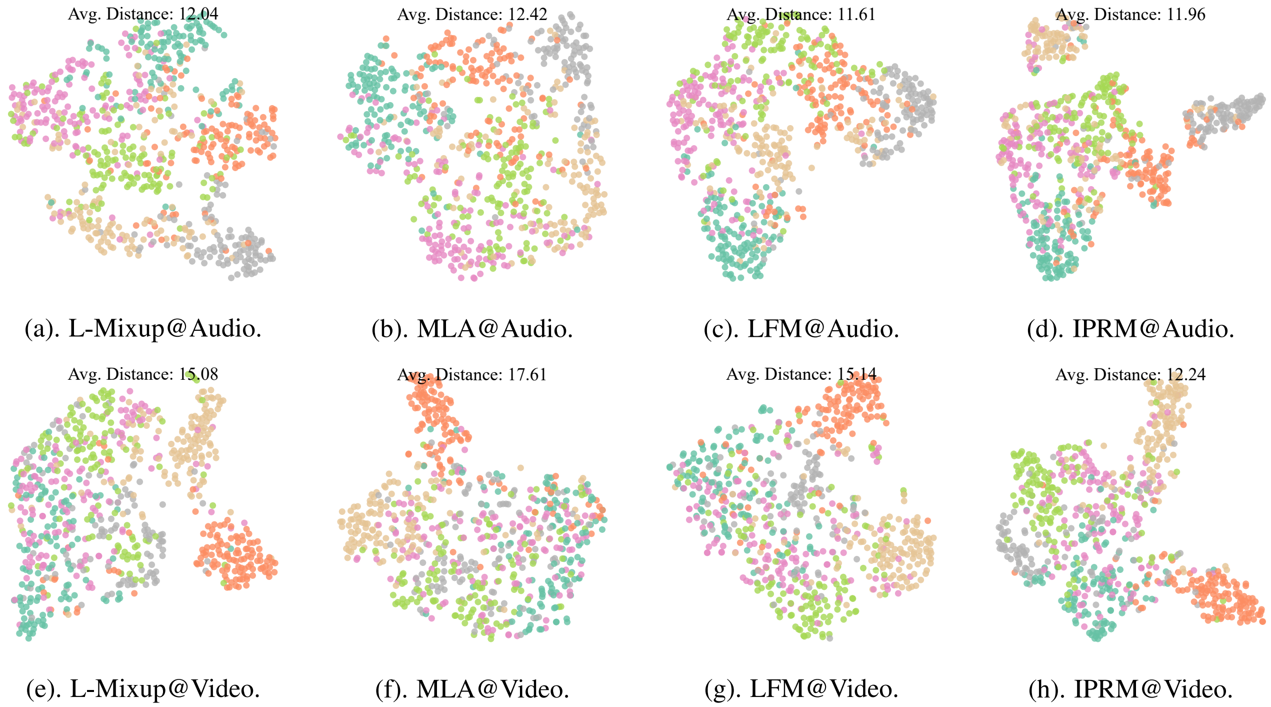
Table 5: Mixup strategy comparison on trimodal dataset.

Computation Cost of Two-Pass Forward: As IPRM adopts a two-pass forward strategy, we compare the computation cost to illustrate the influence of training time. We report the accuracy and computation cost on *CREMA-D* for naive MML, MLA, LFM, and IPRM. The results in Table 4 show that IPRM exhibits slightly slower training compared to naive MML. However, it achieves the best performance among all methods. Notably, the IPRM’s algorithmic complexity remains linear with respect to the sample size N, i.e., $\mathcal{O}(N)$, reflecting its linear scalability. During inference, IPRM follows the standard MML inference procedure and does not introduce any additional computational overhead.

GMM Strategy for Trimodal Dataset: For trimodal datasets, we employ the GMM strategy between any two modalities to achieve modal fusion. Consequently, our structure incorporates three fused classification heads. For instance, in *NVGesture* dataset, the classification heads correspond to RGB and OF, RGB and depth, as well as OF and depth. However, this structure is somewhat cumbersome. Therefore, we explore an alternative strategy to simplify the design of classification heads for trimodal settings.

Specifically, we design random modality sampling strategy for trimodal dataset. In the architecture, we only utilize one classifier for all three modalities. During the forward phase, we randomly select two modalities to perform modality fusion. We compare these two strategies on trimodal datasets, i.e., *NVGesture* and *IEMOCAP* datasets. The results in Table 5 demonstrate that the method based on the random modality sampling strategy achieves comparable results, where the random modality sampling strategy and the strategy adopted in the paper are respectively denoted as “Single-CLS” and “Tri-CLS”. Therefore, in practice, a specific architecture or strategy can be selected based on the requirements.

Fine-Tuning with Pretrained CLIP: We further study the applicability of IPRM for large-scale language vision pretrained models like CLIP [Radford *et al.*, 2021] on *Sarcasm* dataset. The encoders for image and text modalities are replaced by the ViT-B/16 and Transformer from the CLIP



(a). L-Mixup@Audio. (b). MLA@Audio. (c). LFM@Audio. (d). IPRM@Audio.

(e). L-Mixup@Video. (f). MLA@Video. (g). LFM@Video. (h). IPRM@Video.

Figure 4: t-SNE visualization results on *CREMA-D* dataset. The average distance of each multimodal data sample to its respective cluster center is presented in each sub-figure.

Method	Image	Text	Multiple
CLIP	74.82%	82.15%	83.11%
CLIP+MLA	77.45%	83.19%	84.45%
CLIP+LFM	**79.78%**	83.67%	85.42%
CLIP+IPRM	77.46%	**85.43%**	**86.47%**

Table 6: Fine-Tuning with Pretrained CLIP on *Sarcasm* dataset.

model. Then the fine-tuning process is performed to fit the downstream task. The accuracy performance is reported in Table 6, where we utilize "CLIP+*Method*" to denote the specific *Method* to train the models with replaced encoders. According to the results in Table 6, we can see that: (1). Comparing the results of fine-tuning downstream tasks using the rebalanced MML method with those based on the CLIP model reveals that fine-tuning on downstream tasks leads to improved model performance; (2). IPRM can achieve the best accuracy in almost all cases by comparing IPRM with LFM, MLA, and naive CLIP.

Visualization: To study the properties of the learned embeddings for different methods, we present the visualization results by using t-SNE [Van der Maaten and Hinton, 2008] on *CREMA-D* dataset. We select three methods for comparison, including the method with linear mixup (denoted as L-Mixup), MLA, LFM and IPRM.

The t-SNE visualzation results on *CREMA-D* dataset are presented in Figure 4, where we also present the average distance of each multimodal data sample to its respective cluster center in each sub-figure. We can find that our IPRM can achieve smaller distance compared with the remaining approaches, demonstrating that IPRM effectively learns discriminative representations. This effect is particularly pronounced in the video modality, i.e., the weak modality for *CREMA-D* dataset, demonstrating that IPRM more effectively addresses the modality imbalance issue.

6 Conclusion

In this paper, we propose a novel instantaneous probe-and-rebalance multimodal learning approach, namely IPRM, by employing two-forward phase strategy to capture the learning status instantaneously. We utilize geodesic multimodal mixup to probe the strength using dynamic weight. The weights are subsequently recalibrated instantaneously based on the evaluated strength, enabling balanced training through an additional forward pass. This dynamic adjustment is performed continuously throughout the training process. Experiments demonstrate that our IPRM can achieve the best performance on widely used datasets comparing numerous state-of-the-art baselines. In future reseach, related extensions, such as integrating IPRM with attention mechanisms and adapting IPRM to streaming multimodal data scenarios, remain promising for further improving the efficiency and scalability of the approach.

Acknowledgments

This work is supported by the National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

References

[Arandjelovic and Zisserman, 2017] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617. IEEE, 2017.

[Baltrusaitis *et al.*, 2019] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2):423–443, 2019.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *LRE*, 42(4):335–359, 2008.

[Cai *et al.*, 2019] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pages 2506–2515. Association for Computational Linguistics, 2019.

[Cao *et al.*, 2014] 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*, 5(4):377–390, 2014.

[Carreira and Zisserman, 2017] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. Computer Vision Foundation / IEEE, 2017.

[Chang *et al.*, 2015] Angel X. Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. Text to 3d scene generation with rich lexical grounding. In *ACL*, pages 53–62. The Association for Computer Linguistics, 2015.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Du *et al.*, 2022] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Yue Wang, Yang Yuan, and Hang Zhao. Modality laziness: Everybody's business is nobody's business, 2022.

[Fan *et al.*, 2023] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: prototypical modal rebalance for multimodal learning. In *CVPR*, pages 20029–20038. Computer Vision Foundation / IEEE, 2023.

[Fan *et al.*, 2024] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. *CoRR*, abs/2407.19514, 2024.

[Fang *et al.*, 2022] Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. STEMM: self-learning with speech-text manifold mixup for speech translation. In *ACL*, pages 7050–7062. Association for Computational Linguistics, 2022.

[Geng *et al.*, 2022] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *CoRR*, abs/2205.14204, 2022.

[Gong *et al.*, 2023] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *ICLR*. OpenReview.net, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. Computer Vision Foundation / IEEE, 2016.

[Hua *et al.*, 2024] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *ICML*. OpenReview.net, 2024.

[Huang *et al.*, 2021] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *NeurIPS*, pages 10944–10956, 2021.

[Huang *et al.*, 2022] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 9226–9259. PMLR, 2022.

[Li *et al.*, 2023] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pages 22157–22167. IEEE, 2023.

[Molchanov *et al.*, 2016] 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*, pages 4207–4215. Computer Vision Foundation / IEEE, 2016.

[Oh *et al.*, 2023] Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. In *NeurIPS*, 2023.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pages 8228–8237. Computer Vision Foundation / IEEE, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763. PMLR, 2021.

[Sun *et al.*, 2023] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *TPAMI*, 45(3):3200–3225, 2023.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[Verma *et al.*, 2019] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR, 2019.

[Wang *et al.*, 2018] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. FVQA: fact-based visual question answering. *TPAMI*, 40(10):2413–2427, 2018.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12692–12702. Computer Vision Foundation / IEEE, 2020.

[Wei and Hu, 2024] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*. OpenReview.net, 2024.

[Wu *et al.*, 2022] 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*, pages 24043–24055. PMLR, 2022.

[Yang *et al.*, 2015] Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *IJCAI*, pages 1033–1039. AAAI Press, 2015.

[Yang *et al.*, 2019] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *IJCAI*, pages 4092–4098. ijcai.org, 2019.

[Yang *et al.*, 2024a] Yang Yang, Fenqing Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, 2024.

[Yang *et al.*, 2024b] Yang Yang, Wenjuan Xi, Luping Zhou, and Jinhui Tang. Rebalanced vision-language retrieval considering structure-aware distillation. *TIP*, 33:6881–6892, 2024.

[Yang *et al.*, 2025] Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *TPAMI*, 2025.

[Yao and Mihalcea, 2022] Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *ACL*, pages 1824–1834. Association for Computational Linguistics, 2022.

[Yun *et al.*, 2019] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*. IEEE, 2019.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*. OpenReview.net, 2018.

[Zhang *et al.*, 2024] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, pages 27446–27456. IEEE, 2024.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398. IJCAI/AAAI Press, 2016.

[Zhu *et al.*, 2023] Lei Zhu, Tianshi Wang, Fengling Li, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *CoRR*, abs/2308.14263, 2023.

[Zong *et al.*, 2024] Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, and Ken Zheng. Balancing multimodal learning via online logit modulation. In *IJCAI*, pages 5753–5761. ijcai.org, 2024.