Balance-Aware Sequence Sampling Makes Multi-Modal Learning Better

Zhi-Hao Guan , Qing-Yuan Jiang* , Yang Yang*
Nanjing University of Science and Technology
{zhguan, jiangqy, yyang}@njust.edu.cn

Abstract

Multi-modal learning (MML) is frequently hindered by modality imbalance, leading to suboptimal performance in real-world applications. To address this issue, existing approaches primarily focus on rebalancing MML from the perspective of optimization or architecture design. However, almost all existing methods ignore the impact of sample sequences, i.e., an inappropriate training order tends to trigger learning bias in the model, further exacerbating modality imbalance. In this paper, we propose Balance-aware Sequence Sampling (**BSS**) to enhance the robustness of MML. Specifically, we first define a multi-perspective measurer to evaluate the balance degree of each sample in terms of correlation and information criteria. Via this evaluation, we employ a heuristic scheduler based on curriculum learning (CL) that incrementally provides training subsets, progressing from balanced to imbalanced samples to alleviate the imbalance. Moreover, we propose a learning-based probabilistic sampling method to dynamically update the training sequence in a more fine-grained manner, further improving MML performance. Extensive experiments on widely used datasets demonstrate the superiority of our method compared with state-of-the-art (SOTA) baselines. The code is available at https://github.com/njustkmg/IJCAI25-BSS.

1 Introduction

Multi-modal learning has emerged as a prominent research area in artificial intelligence across various scenarios [Yin *et al.*, 2021; Xu *et al.*, 2023; Yang *et al.*, 2021], including speech recognition [Hu *et al.*, 2023], information retrieval [Yang *et al.*, 2024b], and recommender systems [Ye *et al.*, 2025]. By integrating information from diverse sensors, MML has become a driving force in improving performance across these applications. Despite these promising outcomes, MML faces a significant challenge: modality imbalance. Specifically, the inherent heterogeneity of data endows each modality with distinct properties, such as convergence speed [Peng *et al.*, 2022]. As a result, the learning process tends to be dominated by the stronger modality (i.e., the one that converges faster) during joint training, which can lead to insufficient learning of other modalities. In extreme cases, this imbalance may even cause the multi-modal model to perform worse than its best unimodal counterpart [Wang *et al.*, 2020].

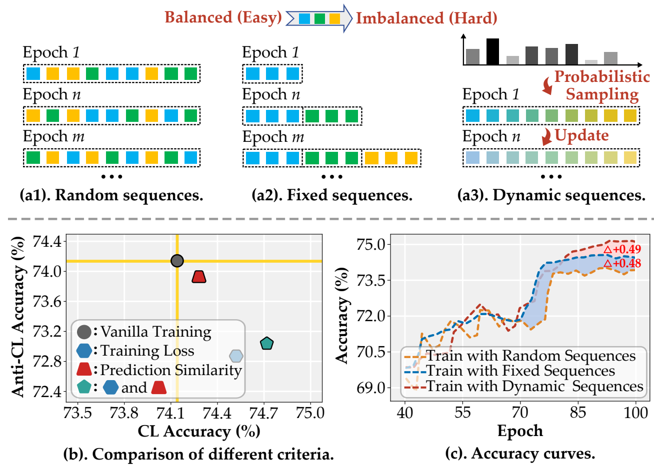

Figure 1: A motivating example of sequence sampling: **(a1-a3).** Illustration of different training sequences. **(b).** Comparison of different criteria under the CL setting. The results show that CL outperforms vanilla training with random sequences, while anti-CL is inferior to it. **(c).** Accuracy curves of different training sequences on the Twitter2015 dataset.

Recently, many impressive studies have been proposed to address the modality imbalance problem from various perspectives [Du *et al.*, 2021; Peng *et al.*, 2022; Li *et al.*, 2023; Yang *et al.*, 2024a; Yang *et al.*, 2025]. Considering the inherent modal differences, a straightforward idea is to manually control the optimization process between strong and weak modalities to achieve rebalancing, such as learning rate adjustment [Yao and Mihalcea, 2022] and gradient modulation [Fan *et al.*, 2023; Peng *et al.*, 2022]. Other approaches attempt to facilitate multi-modal learning through neural architecture design [Du *et al.*, 2021; Xiao *et al.*, 2020]. Although these optimization- and architecture-based methods have shown promising results, they generally overlook an important aspect: *MML can be highly sensitive to the sequence*

*Corresponding authors.

in which training samples are presented at different stages. This motivates us to investigate the role of sample sequences in addressing modality imbalance.

Since the standard training paradigm is characterized by random data shuffling, this process inevitably introduces imbalanced samples into the early training stages (Figure 1 (a1)), which may further exacerbate modality imbalance and ultimately degrade overall performance. To support our viewpoint, we conduct a toy experiment on the Twitter2015 dataset to investigate the relationship between different training sequences and MML performance. Inspired by curriculum learning (CL) [Wang *et al.*, 2022; Soviany *et al.*, 2022], we first evaluate the balance degree of sample pairs based on both correlation criteria (e.g., prediction similarity) and information criteria (e.g., training loss), and then rank them to construct new training sequences (Figure 1 (a2)). The comparison results in Figure 1 (b) reveal an interesting phenomenon: CL effectively boosts MML performance, while anti-CL (i.e., learning from imbalanced to balanced samples) leads to performance degradation across all criteria. This experiment suggests that introducing balanced samples in the early training stages can guide the model toward a more stable and robust optimization path, thereby enhancing overall performance.

Based on our findings, in this paper, we attempt to address the modality imbalance by adjusting the sample sequences, a training paradigm that provides appropriate training samples to the model at different stages. Concretely, we first design a multi-perspective measurer from both correlation and information criteria to evaluate the balance degree of each sample. Via sample evaluation, we propose a heuristic scheduler that progressively constructs training sequences in a balanced-to-imbalanced manner. Moreover, considering that the heuristic scheduler is relatively coarse and may neglect feedback from the current model, we propose a learning-based scheduler that dynamically reconstructs training sequences by assigning sampling probabilities to each data point (Figure 1 (a3)), further enhancing MML performance as shown in Figure 1 (c). To sum up, our contributions are outlined as follows:

- We highlight the critical role of training sequences in addressing modality imbalance, and show that well-structured sequences can significantly improve MML performance.
- We define a multi-perspective measurer to quantify the balance degree of each sample. Based on the resulting balance scores, we then propose both a heuristic and a learning-based sampling method to adjust the training sequences.
- Extensive experiments demonstrate that our proposed method outperforms existing baselines and achieves SOTA performance across widely used datasets.

2 Related Work

2.1 Imbalanced Multi-Modal Learning

Recent research [Peng *et al.*, 2022; Huang *et al.*, 2022] has shown that many multi-modal models fail to outperform the best unimodal counterpart. This phenomenon is attributed to modality imbalance [Fan *et al.*, 2024; Wei *et al.*, 2024b], where each modality cannot be fully learned due to inhibition between them. Considering the existence of both strong and weak modalities, several representative [Wang *et al.*, 2020; Fan *et al.*, 2023; Zong *et al.*, 2024] methods focus on balancing the optimization of individual modalities. In particular, OGM [Peng *et al.*, 2022] introduces an on-the-fly gradient modulation technique, which adaptively adjusts the optimization process for each modality by monitoring the discrepancy in their contributions to the learning objective. PMR [Fan *et al.*, 2023] uses prototypes to control the update direction for improved unimodal performance. Other studies [Du *et al.*, 2021; Wu *et al.*, 2022] attempt to boost MML performance by introducing supplementary modules. For instance, UMT [Du *et al.*, 2021] trains the multi-modal model with knowledge distillation [Gou *et al.*, 2021] from well-learned teacher encoders to obtain richer unimodal representations. However, these methods increase model complexity and training costs. In this paper, from the perspective of sample sequences, we address modality imbalance by guiding the model to progressively learn training samples in a balanced-to-imbalanced manner, without the need for additional modules.

2.2 Sequence-oriented Multi-modal Learning

Sequence-oriented MML is crucial in machine learning, enabling models to train on a meaningful subset derived from the original dataset distribution. These strategies are primarily applied in two areas: curriculum learning (CL) [Soviany *et al.*, 2022] and noisy label learning (NLL) [Patel and Sastry, 2023]. CL is a training paradigm that progresses from easier samples to harder ones. By guiding the model toward a better parameter space, CL has been widely adopted across various fields, including large language models [Wang *et al.*, 2024], action recognition [Tong *et al.*, 2023], and reinforcement learning [Narvekar *et al.*, 2020]. A typical curriculum system consists of two main components: a difficulty measurer to evaluate the learning difficulty of samples and a scheduler to manage the assignment of training subsets. On the other hand, sequence-oriented MML in NLL focuses on selecting clean samples from a noisy training set for model learning. Commonly used criteria for identifying noisy labels, such as training loss [Wei *et al.*, 2020], Jensen-Shannon divergence [Xu *et al.*, 2025], and representation similarity [Ortego *et al.*, 2021], facilitate reliable data selection and ultimately enhance model robustness. Inspired by the core idea of sequence-oriented MML, we prioritize balanced samples to address modality imbalance. This strategy helps rebalance the training process, enabling our method to learn robust feature representations while avoiding early-stage optimization dilemmas.

3 Methodology

In this section, we present our proposed method in detail. The overall architecture is shown in Figure 2, which consists of two main components: a multi-modal training framework for learning representations, and a Balance-aware Sequence Sampling (BSS) module for rebalancing MML via a multi-perspective measurer and two optional schedulers (one heuristic and the other learning-based).

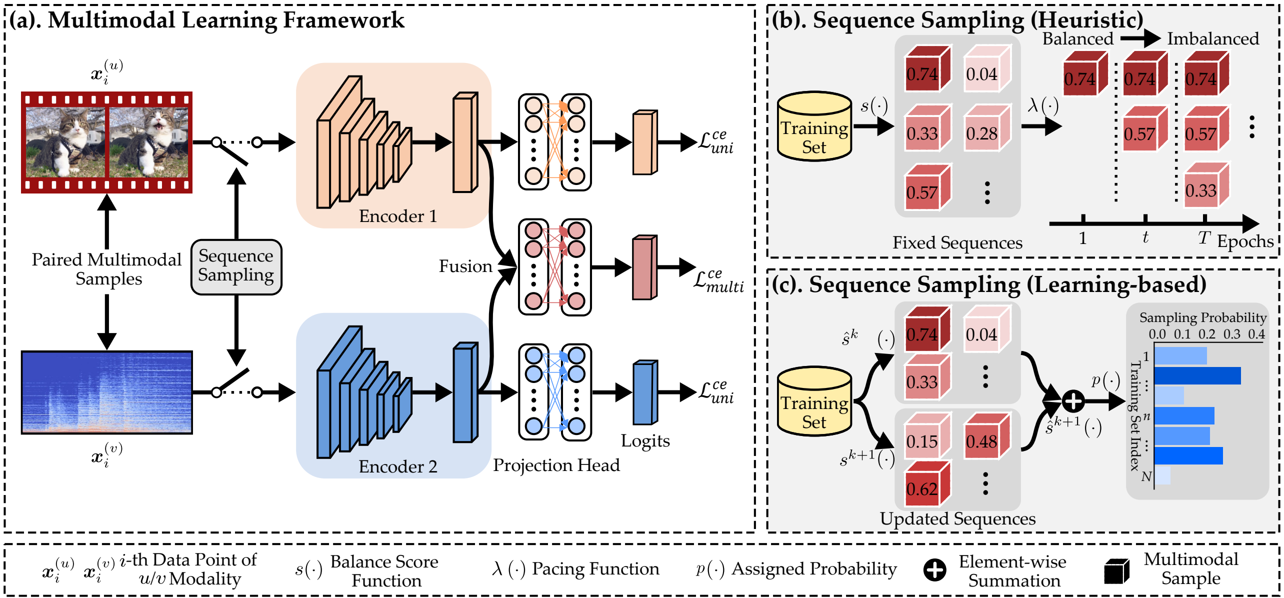

Figure 2: Illustration of our method: **(a).** Multi-modal learning framework. **(b).** Sequence sampling with a heuristic scheduler. **(c).** Sequence sampling with a learning-based scheduler.

3.1 Preliminary

Without loss of generality, we consider a multi-modal sample with u and v modalities. Formally, let $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ denote the dataset, where $\boldsymbol{X} = \{\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_i^{(v)}\}_{i=1}^n$ represents n training samples and $\boldsymbol{Y} = \{\boldsymbol{y}_i \,|\, \boldsymbol{y}_i \in \{0,1\}^c\}_{i=1}^n$ is the corresponding category labels with a total of c categories. MML aims to train a model to predict the category label of a given multi-modal sample.

For the MML prediction task, we typically employ deep neural networks to learn feature embeddings of each modality from the original space. We use $f^{(j)}(\cdot)$ as the feature extractor for j-th modality, $j \in \{u, v\}$. Given a sample $\boldsymbol{x}_i^{(j)}$, the feature extraction can be expressed as:

$$\boldsymbol{e}_i^{(j)} = f^{(j)}(\boldsymbol{x}_i^{(j)}; \theta^{(j)}), \tag{1}$$

where $\boldsymbol{e}_i^{(j)} \in \mathbb{R}^d$ denotes the d-dimension feature embedding of $\boldsymbol{x}_i^{(j)}$, the $\theta^{(j)}$ denotes the learnable parameters of j-th encoder. After extracting feature embeddings for all modalities, we adopt a fusion function $g(\cdot)$ to fuse them, followed by a classifier (e.g., a fully connected layer) to map the feature embedding to $\mathbb{R}^c$. This procedure can be formulated as:

$$\boldsymbol{e}_i = g(\boldsymbol{e}_i^{(u)}, \boldsymbol{e}_i^{(v)}), \quad \hat{\boldsymbol{y}}_i = \texttt{softmax}(\boldsymbol{W}\boldsymbol{e}_i + \boldsymbol{b}). \tag{2}$$

Here, $\boldsymbol{W} \in \mathbb{R}^{c \times D}$, $\boldsymbol{b} \in \mathbb{R}^c$ denote the weights and bias of the classifier, respectively, and D denotes the dimension of $\boldsymbol{e}_i$. Then, the objective function of multi-modal learning can be formulated as:

$$\ell_{multi}^{ce}(\boldsymbol{X}, \boldsymbol{Y}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i^\top \log \hat{\boldsymbol{y}}_i. \tag{3}$$

Considering that MML can benefit from the supervision of unimodal predictions [Zou *et al.*, 2023], we jointly optimize both multi-modal and unimodal objectives. Therefore, the final objective function can be reformulated as:

$$\ell_{total} = (1-\alpha)\ell_{multi}^{ce}(\boldsymbol{X}, \boldsymbol{Y}) + \alpha \sum_{j \in \{u,v\}} \ell_{uni}^{ce}(\boldsymbol{X}^{(j)}, \boldsymbol{Y}), \tag{4}$$

where α denotes the weighted parameter between the losses.

3.2 Multi-perspective Measurer

To construct well-structured training sequences that address modality imbalance, we next introduce how to measure the balance degree of a multi-modal sample from the perspectives of correlation and information criteria.

Correlation Criterion: Different modalities exhibit inherent cross-modal correlation, as they describe the same concept via diverse representations, capturing complementary information. Although cross-modal correlation can be measured from various aspects, we focus on the most commonly used and critical criterion, i.e., prediction similarity. The reason is that prediction similarity directly measures output consistency. Formally, given a sample $\boldsymbol{x}_i = \{\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_i^{(v)}\}$, the prediction similarity is defined as:

$$\texttt{sim}(\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_i^{(v)}) = \frac{[\hat{\boldsymbol{y}}_i^{(u)}]^\top \hat{\boldsymbol{y}}_i^{(v)}}{\|\hat{\boldsymbol{y}}_i^{(u)}\|_2 \|\hat{\boldsymbol{y}}_i^{(v)}\|_2}. \tag{5}$$

Here, $\|\cdot\|_2$ denotes L_2 norm of the unimodal predictions.

Information Criterion: While prediction similarity reflects the balance between modalities, it does not verify whether the predictions of each modality are correct. In other words, high prediction similarity may still occur even when all modalities produce incorrect predictions. Therefore, we further introduce label-related training loss as an intuitive metric to evaluate the learning difficulty of each sample.

Balance Score: We denote the sets of prediction similarity and training loss computed over all n training samples as $\mathcal{S} = [\mathtt{sim}(\boldsymbol{x}_1^{(u)}, \boldsymbol{x}_1^{(v)}), \ldots, \mathtt{sim}(\boldsymbol{x}_n^{(u)}, \boldsymbol{x}_n^{(v)})]$ and $\mathcal{L} = [\ell_{total}(\boldsymbol{x}_1^{(u)}, \boldsymbol{x}_1^{(v)}, \boldsymbol{y}_1), \ldots, \ell_{total}(\boldsymbol{x}_n^{(u)}, \boldsymbol{x}_n^{(v)}, \boldsymbol{y}_n)]$, respectively. Hence, the balance score of sample $\boldsymbol{x}_i = \{\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_i^{(v)}\}$ can be formulated as the combination of correlation criterion and information criterion:

$$\begin{aligned} s(\boldsymbol{x}_i) = & \frac{\mathtt{sim}(\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_i^{(v)}) - \mathtt{min}(\mathcal{S})}{\max(\mathcal{S}) - \mathtt{min}(\mathcal{S})} \\ & - \frac{\ell_{total}(\boldsymbol{x}_i^{(u)}, \boldsymbol{x}_i^{(v)}, \boldsymbol{y}_i) - \mathtt{min}(\mathcal{L})}{\max(\mathcal{L}) - \mathtt{min}(\mathcal{L})}. \end{aligned} \tag{6}$$

This normalization ensures that both terms lie on a comparable scale. A higher score $s(\boldsymbol{x}_i) \in [-1, 1]$ indicates a better balance degree.

3.3 Training Scheduler

After evaluating the balance score of each sample, we proceed to control the presentation order of training data from balanced to imbalanced samples, i.e., the sample sequence for each training epoch.

Similar to human education, if teachers impart knowledge all at once, students may become overwhelmed and fail to learn effectively. On the other hand, if teachers focus too much on basic knowledge, students may lose motivation. In the following, we introduce a coarse but efficient heuristic scheduler and a more refined, effective learning-based scheduler for constructing sample sequences.

Heuristic Scheduler: Inspired by curriculum learning [Wang *et al.*, 2022; Soviany *et al.*, 2022], we rank the training samples from balanced to imbalanced according to the defined balance score, and then employ a pace function [Hacohen and Weinshall, 2019] to determine the number of samples included in the training set at each epoch. In practice, various pacing functions exist, such as the baby step [Bengio *et al.*, 2009], linear function [Wang *et al.*, 2022], and root function [Platanios *et al.*, 2019]. However, the impact of existing pacing functions on modality imbalance is not the focus of our work. Here, we adopt a widely used root function $\lambda(t)$ to achieve this:

$$\lambda(t) = \mathtt{min}\left(1, \sqrt{\frac{1-\lambda_0^2}{T_{grow}} \cdot t + \lambda_0^2}\right), \tag{7}$$

where T_{grow} represents the training epoch when this function first reaches 1, and $\lambda_0 \in (0, 1]$ is the initial proportion of the training samples. $\lambda(t)$ maps the training epoch t to an interval $\lambda \in (0, 1]$, which means λ proportion of the most balanced samples are available at t-th epoch. This function starts at $\lambda(0) > 0$ and ends at $\lambda(T_{grow}) = 1$.

From Equation 7, the pace function serves as a threshold that progressively expands the sampling space during training. At each epoch t, the current batch data $\boldsymbol{X}_{batch}$ is randomly sampled from the top λ proportion of training data in the entire ranked sequence $\boldsymbol{X}_{rank}$:

$$\boldsymbol{X}_{batch}(t) = \mathtt{Sampling}\left(\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \boldsymbol{X}_{rank}, i < \lfloor n \cdot \lambda(t) \rfloor\}\right), \tag{8}$$

where n denotes the number of training samples. Thus, the heuristic scheduler allows the model to focus on balanced samples during the early training stages and gradually broaden the learning scope by incorporating those imbalanced ones. Please note that the sample evaluation is performed only once before model training, which means $\boldsymbol{X}_{rank}$ is a fixed sequence.

Learning-based Scheduler: Despite the simplicity and efficiency of the heuristic scheduler in practice, it has one main limitation: the fixed training sequence is coarse-grained, which may neglect feedback from the current model and potentially lead to inaccurate sample evaluation. Therefore, we further propose a learning-based scheduler that flexibly addresses the above limitation. This scheduler reconstructs the dynamic sequence by learning a sampling probability for each data point, considering both the balance of past and current samples in a more fine-grained manner.

Specifically, we refer to the balance score $s(\boldsymbol{x}_i)$ in Equation 6 and update it in a certain epoch interval, using E for short. Subsequently, the $k+1$-th balance score can be expressed as:

$$\hat{s}^{k+1}(\boldsymbol{x}_i) = \begin{cases} s^{k+1}(\boldsymbol{x}_i), & \text{if } k = 0, \\ (1-\beta)\hat{s}^k(\boldsymbol{x}_i) + \beta s^{k+1}(\boldsymbol{x}_i), & \text{otherwise,} \end{cases} \tag{9}$$

where $k = \lfloor t/E \rfloor$, t denotes the t-th epoch, β is an adjustment parameter, and s^1 is the balance score obtained before model training, i.e., the initial evaluation results.

According to the updated balance scores, the learned sampling probability p for each data point $\boldsymbol{x}_i$ in the t-th epoch can be computed using the softmax operation:

$$p(\boldsymbol{x}_i) = \frac{e^{\hat{s}^{k+1}(\boldsymbol{x}_i)}}{\sum_{j=1}^{n} e^{\hat{s}^{k+1}(\boldsymbol{x}_j)}}. \tag{10}$$

Finally, in the t-th epoch, each data point $\boldsymbol{x}_i$ is sampled with probability $p(\boldsymbol{x}_i)$ to construct the current batch data $\boldsymbol{X}_{batch}$, without replacement. This process is formulated as:

$$\boldsymbol{X}_{batch}(t) = \mathtt{Sampling}(\{p(\boldsymbol{x}_1), p(\boldsymbol{x}_2), \ldots, p(\boldsymbol{x}_n)\}). \tag{11}$$

Hence, training data with higher sampling probabilities (i.e., more balanced ones) are preferentially selected for the mini-batch in each epoch.

Discussion: Our proposed method aims to address the modality imbalance problem through sequence sampling in a balanced-to-imbalanced learning manner. Thus, our BSS can be integrated as a model-independent plugin into most existing MML approaches.

3.4 Model Inference

After training, the learned model can be applied for prediction during the inference stage. Following [Fan *et al.*, 2024; Zhang *et al.*, 2024], we adopt a simple weighted combination of logits output from each modality and their fusion, represented as $\boldsymbol{z}_{total} = \boldsymbol{z}_{multi} + \sum_{j\in\{u,v\}} \boldsymbol{z}_{uni}^{(j)}$. Subsequently,

Algorithm 1: Multi-modal Learning with Balance-aware Sequence Sampling (BSS).

Input : Training set $\boldsymbol{X}_{train}$, category labels $\boldsymbol{Y}_{train}$.
Output: Learned parameters θ of all models.
INIT Initialize parameters θ^0, maximum epochs T, training set for ranking $\boldsymbol{X}_{rank} = \emptyset$, curriculum period T_{grow}, initial proportion λ_0, epoch interval E.

```
/* Calculate the balance score via measurer. */
for each sample x_i in X_train do
    Obtain balance score s(x_i) based on Equation 6.
    Add x_i to X_rank in descending order of s(x_i).
end
/* Train model using sample sequences from scheduler. */
for t = 0 to T − 1 do
    if scheduler == 'heuristic' then
        Calculate the proportion of the training samples
          λ(t) with Equation 7.
        Obtain current batch data X_batch from X_rank with
          Equation 8.
    else if scheduler == 'learning-based' then
        Update s(x_i) every E epochs with Equation 9.
        Assign sampling probability p(x_i) based on s(x_i)
          with Equation 10.
        Obtain current batch data X_batch with Equation 11.
    end
    Train model with X_batch and update parameters θ.
    Update t = t + 1.
end
```

the predicted category $\hat{y}$ for a given unseen multi-modal sample can be denoted as:

$$\hat{y} = \underset{i}{\texttt{argmax}} \frac{e^{\boldsymbol{z}^i_{total}}}{\sum_{j=1}^{c} e^{\boldsymbol{z}^j_{total}}}. \tag{12}$$

4 Experiments

4.1 Experimental Setup

Datasets: We validate our proposed method on six widely used datasets, including CREMA-D [Cao *et al.*, 2014], Kinetics-Sounds [Arandjelovic and Zisserman, 2017], VGGSound [Chen *et al.*, 2020], Twitter2015 [Yu and Jiang, 2019], Sarcasm [Cai *et al.*, 2019], and NVGesture [Molchanov *et al.*, 2016]. Among them, CREMA-D, Kinetics-Sounds, and VGGSound contain both audio and video modalities. CREMA-D includes 7,442 video clips across six emotional categories, with 6,698 clips for training and 744 for testing. Kinetics-Sounds is categorized into 31 distinct actions, split into 15,000 for training, 1,900 for validation, and 1,900 for testing. VGGSound provides 168,618 videos for training and validation, along with 13,954 videos for testing. Moreover, Twitter2015 and Sarcasm datasets involve both image and text modalities. Twitter2015 comprises 5,338 text-image pairs, divided into 3,179 for training, 1,122 for validation, and 1,037 for testing. Sarcasm contains 24,635 text-image pairs, allocated as 19,816 for training, 2,410 for validation, and 2,409 for testing. Lastly, NVGesture features three modalities, i.e., RGB, optical flow (OF), and Depth, with 1,050 samples for training and 482 samples for testing.

Baselines and Evaluation Metrics: We conduct a comprehensive comparison of BSS with two types of baselines: vanilla fusion methods and multi-modal rebalance approaches. The former includes Concat, Affine [Perez *et al.*, 2018], Channel, ML-LSTM [Nie *et al.*, 2021], Sum, Weight, and ETMC [Han *et al.*, 2023]. The latter comprises MSES [Fujimori *et al.*, 2019], OGR-GB [Wang *et al.*, 2020], DOMFN [Yang *et al.*, 2022], OGM [Peng *et al.*, 2022], MSLR [Yao and Mihalcea, 2022], AGM [Li *et al.*, 2023], PMR [Fan *et al.*, 2023], ReconBoost [Hua *et al.*, 2024], MMPareto [Wei and Hu, 2024], SMV [Wei *et al.*, 2024a], MLA [Zhang *et al.*, 2024], and AMSS [Yang *et al.*, 2025].

Following [Peng *et al.*, 2022; Hua *et al.*, 2024], we utilize accuracy (ACC), mean average precision (MAP), and Macro F1-score (Mac-F1) as evaluation metrics. ACC measures the ratio of correct predictions to total predictions. MAP reflects the average precision across all samples, while Mac-F1 computes the average of F1 scores across all categories.

Implementation Details: Following [Fan *et al.*, 2023], for audio-video datasets, we use ResNet18 [He *et al.*, 2016] as the backbone to encode each modality. For text-image datasets, we employ ResNet50 for images and BERT [Devlin *et al.*, 2019] for text processing. For the trimodal dataset NVGesture, we follow the setup of [Wu *et al.*, 2022] and adopt the I3D [Carreira and Zisserman, 2017] as the unimodal backbone. To ensure fairness, all methods use the same backbone during training. The optimizer for audio-video datasets is stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 10^{-4}. The initial learning rate is set to 10^{-2} and is reduced by a factor of 10 when the loss saturates. The batch size is set to 64 for CREMA-D and Kinetics-Sounds, 16 for VGGSound, and 2 for NVGesture. For text-image datasets, we employ the Adam optimizer starting with a learning rate of 10^{-5}, with a batch size of 64. Furthermore, the hyperparameters α and β are set to 0.2 and 0.6, respectively. For the training scheduler, the curriculum period T_{grow} and the initial proportion λ_0 are configured as 40 and 0.1 under the heuristic setting, while the epoch interval E is configured as 5 under the learning-based setting. All models are trained on an NVIDIA GeForce RTX 3090 GPU.

4.2 Comparison with SOTA MML Baselines

We conduct comprehensive comparisons to assess the superiority of our proposed method in addressing the imbalanced MML problem. The classification performance across all datasets is reported in Table 1 and Table 2, where "BSS-H" and "BSS-L" denote the proposed method with the heuristic scheduler and learning-based scheduler, respectively. Please note that "-" in Table 1 denotes that the corresponding methods are not applicable to the respective datasets.

Results on Bimodal Dataset: Referring to the first four datasets in Table 1, we derive the following key observations: (1). Unimodal performance may outperform multi-modal joint training. For instance, the text-modal performance on the Twitter2015 dataset is obviously better than most vanilla fusion methods, indicating an inhibitory rela-

Method	CREMA-D		Kinetics-Sounds		Twitter2015		Sarcasm		NVGesture	
	ACC (%)	MAP (%)	ACC (%)	MAP (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)
Audio/Text/RGB	63.17	68.61	54.12	56.69	73.67	68.49	81.36	80.65	78.22	78.33
Video/Image/OF	45.83	58.79	55.62	58.37	58.63	43.33	71.81	70.73	78.63	78.65
Depth	-	-	-	-	-	-	-	-	81.54	81.83
Concat	63.31	68.31	64.55	71.31	70.11	63.86	82.86	82.43	81.33	81.47
Affine	66.26	71.93	64.24	69.31	72.03	59.92	82.47	81.88	82.78	82.81
Channel	66.13	71.75	63.51	68.66	-	-	-	-	81.54	81.57
ML-LSTM	62.94	64.73	63.84	69.02	70.68	65.64	82.05	70.73	83.20	83.30
Sum	63.44	69.08	64.97	71.03	73.12	66.61	82.94	82.47	82.99	83.05
Weight	66.53	73.26	65.33	71.33	72.42	65.16	82.65	82.19	83.42	83.57
ETMC	65.86	71.34	65.67	71.19	73.96	67.39	83.69	83.23	83.61	83.69
MSES	61.56	68.83	64.71	70.63	71.84	66.55	84.18	83.60	81.12	81.47
OGR-GB	64.65	84.54	67.10	71.39	74.35	68.69	83.35	82.71	82.99	83.05
DOMFN	67.34	85.72	66.25	72.44	74.45	68.57	83.56	82.62	-	-
OGM	66.94	71.73	66.06	71.44	74.92	68.74	83.23	82.66	-	-
MSLR	65.46	71.38	65.91	71.96	72.52	64.39	84.23	83.69	82.86	82.92
AGM	67.07	73.58	66.02	72.52	74.83	69.11	84.02	83.44	82.78	82.82
PMR	66.59	70.30	66.56	71.93	74.25	68.60	83.60	82.49	-	-
ReconBoost	74.84	81.24	70.85	74.24	74.42	68.34	84.37	83.17	84.13	<u>86.32</u>
MMPareto	74.87	85.35	70.00	78.50	73.58	67.29	83.48	82.48	83.82	84.24
SMV	78.72	84.17	69.00	74.26	74.28	68.17	84.18	83.68	83.52	83.41
MLA	79.43	85.72	70.04	74.13	73.52	67.13	84.26	83.48	83.40	83.72
AMSS	70.30	76.14	72.25	<u>79.13</u>	<u>75.12</u>	<u>69.23</u>	84.35	83.77	84.64	84.94
BSS-H	<u>80.78</u>	<u>87.86</u>	<u>72.67</u>	78.61	74.73	68.67	<u>84.41</u>	<u>83.86</u>	<u>85.06</u>	85.15
BSS-L	**82.80**	**88.61**	**73.95**	**79.43**	**75.22**	**69.51**	**85.01**	**84.62**	**86.72**	**87.04**

Table 1: Comparison with SOTA multi-modal learning methods. The best performances are highlighted in bold, and the second best is underlined. Higher ACC, MAP, or F1 scores indicate better performance.

tionship between different modalities. (2). Most multi-modal rebalance approaches demonstrate significant improvements over vanilla fusion methods. This phenomenon not only underscores the adverse impact of modality imbalance on performance but also validates the effectiveness of the multi-modal rebalance approach. (3). Compared to all baselines, including both vanilla fusion methods and multi-modal rebalance approaches, our proposed method achieves the best performance by a large margin across all metrics. It can be observed that BSS-L delivers significant performance improvements on both the CREMA-D and Kinetics-Sounds datasets. After sequence sampling, our method surpasses the best baseline (MLA) [Zhang *et al.*, 2024] with gains of 3.37%/2.89% and 3.91%/5.30% in ACC and MAP metrics, respectively.

Results on Trimodal Dataset: In addition, we present a comparison with SOTA baselines on the NVGesture dataset. As shown in the last dataset of Table 1, unlike multi-modal rebalance approaches limited to scenarios with only two modalities (e.g., OGM [Peng *et al.*, 2022] and PMR [Fan *et al.*, 2023]), our method effectively tackles the challenges in scenarios involving more than two modalities and achieves the best performance.

Results on Large-scale Dataset: To further evaluate the generality of our method, we conduct experiments on the large-scale VGGSound dataset. Given the size of the dataset, we select a few representative baselines for comparison, including OGM, AGM [Li *et al.*, 2023], ReconBoost [Hua *et al.*, 2024], MMPareto [Wei and Hu, 2024], SMV [Wei *et al.*, 2024a], and MLA. The results in Table 2 consistently demonstrate that our BSS-L achieves superior performance.

Method	ACC (%)	MAP (%)
OGM	48.29	49.78
AGM	47.11	51.98
ReconBoost	50.97	53.87
MMPareto	51.25	54.73
SMV	50.31	53.62
MLA	<u>51.65</u>	54.73
BSS-H	51.61	<u>55.68</u>
BSS-L	**52.80**	**56.61**

Table 2: Performances on the VGGSound dataset.

4.3 Ablation Study

We conduct ablation studies to verify the effectiveness of using different criteria for sample evaluation, namely unimodal prediction similarity (PreSim) and training loss (Loss). Table 3 presents the results under the learning-based setting, which reveal that: (1). Vanilla training may exacerbate modality imbalance. For instance, when the video modality converges, the audio modality remains insufficiently trained, leading to a significant gap between the two modalities (4.66%/6.41% in ACC/MAP). (2). Both "PreSim" and "Loss", when employed, can boost classification performance. (3). By integrating "PreSim" and "Loss", BSS-L achieves the best performance. This is predictable, as prioritizing balanced samples based on correlation and information criteria helps narrow the gap between modalities, facilitating both unimodal and multi-modal learning processes.

Criterion		ACC (%) / MAP (%)		
PreSim	Loss	Audio	Video	Multi
✗	✗	49.37/51.07	54.03/57.48	70.44/76.62
✗	✓	52.11/54.40	54.23/57.91	72.44/79.41
✓	✗	52.38/54.32	**54.93/58.52**	73.25/78.98
✓	✓	**52.73/54.43**	54.74/58.46	**73.95/79.43**

Table 3: Ablation study on the Kinetics-Sounds dataset under the learning-based setting.

4.4 Further Analysis

Sensitivity to Hyperparameters: In calibrating our proposed method, we identify two hyperparameters: α in Equation 4 and β in Equation 9, determining the strength for balancing classification loss and regulating the balance score, respectively. Figure 3 (a) depicts the performance of different α. As α increases, the accuracy of our method first increases and then decreases. This shows that proper unimodal learning has a promoting effect, but over-considering unimodal optimization may hinder multi-modal interactions. From Figure 3 (b), we can find that the performance is marginally affected by β, highlighting the insensitivity of our method to hyperparameters. Despite some fluctuations, our method still demonstrates excellent effectiveness, i.e., being consistently better than baseline vanilla MML.

Robustness of the Pre-trained Model: We further explore the robustness of the large pre-trained model on text-image datasets. We replace each modality encoder with the corresponding encoder pre-trained by CLIP [Radford *et al.*, 2021] and fine-tune the model. The results are shown in Figures 3 (c) and (d), where "CLIP+MLA" and "CLIP+Ours" represent the use of MLA and our approach, respectively. From the results, we can draw the following observations: (1). Both "CLIP+MLA" and "CLIP+Ours" outperform CLIP in all cases. (2). Via sequence sampling, our method achieves better performance than MLA.

Case Study: We investigate whether our method can effectively distinguish between balanced and imbalanced samples in a randomly ordered sequence. From the representative samples in Figure 4, we observe that balanced samples exhibit strong semantic consistency between modalities, as indicated by high balance scores, while imbalanced samples typically display weak semantic connections or irrelevant information.

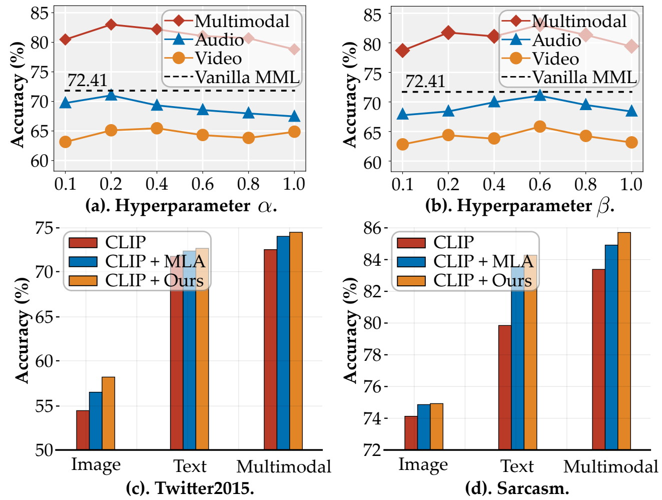

Figure 3: **(a).** and **(b).** Sensitivity to hyperparameters α and β on the CREMA-D dataset. **(c).** and **(d).** Robust performance achieved by using the CLIP pre-trained model as encoders.

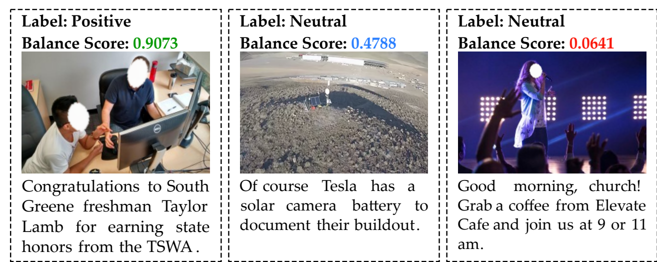

Figure 4: Qualitative results of sample evaluation. We present some representative samples selected from different segments (early, middle, and late) of the training sequence after evaluation under the heuristic setting.

5 Conclusion

In this paper, we propose a novel multi-modal learning method called Balance-aware Sequence Sampling (BSS). By defining a multi-perspective measurer, we evaluate the balance score of each sample. Via this evaluation, we design a heuristic and a learning-based scheduler to construct sample sequences for the model at different training stages. As a result, BSS addresses modality imbalance through a balanced-to-imbalanced learning strategy, thereby enhancing MML performance. Furthermore, BSS can be integrated as a model-independent plugin into most existing MML approaches. Extensive experiments on widely used datasets demonstrate the superiority of BSS over SOTA baselines.

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), National Key Laboratory of Information Systems Engineering (05202403), Fundamental Research Funds for the Central Universities (No. 30925010205).

References

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

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, volume 382, pages 41–48, 2009.

[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, 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, 2017.

[Chen *et al.*, 2020] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725, 2020.

[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-HLT*, pages 4171–4186, 2019.

[Du *et al.*, 2021] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *CoRR*, abs/2106.11059, 2021.

[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, 2023.

[Fan *et al.*, 2024] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. In *ACMMM*, pages 5470–5478, 2024.

[Fujimori *et al.*, 2019] Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-specific learning rate control for multimodal classification. In *ACPR*, volume 12047, pages 412–422, 2019.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129(6):1789–1819, 2021.

[Hacohen and Weinshall, 2019] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *ICML*, volume 97, pages 2535–2544, 2019.

[Han *et al.*, 2023] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *TPAMI*, 45(2):2551–2566, 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, 2016.

[Hu *et al.*, 2023] Yuchen Hu, Ruizhe Li, Chen Chen, Heqing Zou, Qiushi Zhu, and Eng Siong Chng. Cross-modal global interaction and local alignment for audio-visual speech recognition. In *IJCAI*, pages 5076–5084, 2023.

[Hua *et al.*, 2024] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *ICML*, pages 1–25, 2024.

[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, pages 9226–9259, 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, 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, 2016.

[Narvekar *et al.*, 2020] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *JMLR*, 21:181:1–181:50, 2020.

[Nie *et al.*, 2021] Weizhi Nie, Yan Yan, Dan Song, and Kun Wang. Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. *MTA*, 80(11):16205–16214, 2021.

[Ortego *et al.*, 2021] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *CVPR*, pages 6606–6615, 2021.

[Patel and Sastry, 2023] Deep Patel and P. S. Sastry. Adaptive sample selection for robust learning under label noise. In *WACV*, pages 3921–3931, 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, 2022.

[Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951, 2018.

[Platanios *et al.*, 2019] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. Competence-based curriculum learning for neural machine translation. In *NAACL-HLT*, pages 1162–1172, 2019.

[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, 2021.

[Soviany *et al.*, 2022] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *IJCV*, 130(6):1526–1565, 2022.

[Tong *et al.*, 2023] Anyang Tong, Chao Tang, and Wenjian Wang. Semi-supervised action recognition from temporal augmentation using curriculum learning. *TCSVT*, 33(3):1305–1319, 2023.

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

[Wang *et al.*, 2022] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *TPAMI*, 44(9):4555–4576, 2022.

[Wang *et al.*, 2024] Xin Wang, Yuwei Zhou, Hong Chen, and Wenwu Zhu. Curriculum learning for multimedia in the era of large language models. In *ACMMM*, pages 11296–11297, 2024.

[Wei and Hu, 2024] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, pages 1–14, 2024.

[Wei *et al.*, 2020] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13723–13732, 2020.

[Wei *et al.*, 2024a] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pages 27328–27337, 2024.

[Wei *et al.*, 2024b] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *ECCV*, volume 15122, pages 71–86, 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*, volume 162, pages 24043–24055, 2022.

[Xiao *et al.*, 2020] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *CoRR*, abs/2001.08740, 2020.

[Xu *et al.*, 2023] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *TPAMI*, 45(10):12113–12132, 2023.

[Xu *et al.*, 2025] Yuanzhuo Xu, Xiaoguang Niu, Jie Yang, Ruiyi Su, Jian Zhang, Shubo Liu, and Steve Drew. Revisiting interpolation for noisy label correction. In *AAAI*, pages 21833–21841, 2025.

[Yang *et al.*, 2021] Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, and Yuan Jiang. Semi-supervised multi-modal clustering and classification with incomplete modalities. *TKDE*, 33(2):682–695, 2021.

[Yang *et al.*, 2022] Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu. DOMFN: A divergence-orientated multi-modal fusion network for resume assessment. In *ACMMM*, pages 1612–1620, 2022.

[Yang *et al.*, 2024a] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, pages 62108–62122, 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*, pages 1–14, 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, 2022.

[Ye *et al.*, 2025] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. Harnessing multimodal large language models for multimodal sequential recommendation. In *AAAI*, pages 13069–13077, 2025.

[Yin *et al.*, 2021] Ziyi Yin, Ruijin Liu, Zhiliang Xiong, and Zejian Yuan. Multimodal transformer networks for pedestrian trajectory prediction. In *IJCAI*, pages 1259–1265, 2021.

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

[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, 2024.

[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, 2024.

[Zou *et al.*, 2023] Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. Unis-mmc: Multimodal classification via unimodality-supervised multimodal contrastive learning. In *ACL*, pages 659–672, 2023.