Learning to Rebalance Multi-Modal Optimization by Adaptively Masking Subnetworks

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang

Abstract—Multi-modal learning aims to enhance performance by unifying models from various modalities but often faces the “modality imbalance” problem in real data, leading to a bias towards dominant modalities and neglecting others, thereby limiting its overall effectiveness. To address this challenge, the core idea is to balance the optimization of each modality to achieve a joint optimum. Existing approaches often employ a modal-level control mechanism for adjusting the update of each modal parameter. However, such a global-wise updating mechanism ignores the different importance of each parameter. Inspired by subnetwork optimization, we explore a uniform sampling-based optimization strategy and find it more effective than global-wise updating. According to the findings, we further propose a novel importance sampling-based, element-wise joint optimization method, called Adaptively Mask Subnetworks Considering Modal Significance (AMSS). Specifically, we incorporate mutual information rates to determine the modal significance and employ non-uniform adaptive sampling to select foreground subnetworks from each modality for parameter updates, thereby rebalancing multi-modal learning. Additionally, we demonstrate the reliability of the AMSS strategy through convergence analysis. Building upon theoretical insights, we further enhance the multi-modal mask subnetwork strategy using unbiased estimation, referred to as AMSS+. Extensive experiments reveal the superiority of our approach over comparison methods.

Index Terms—Multi-Modal Learning, Modality Imbalance, Subnetwork Optimization

1 Introduction

In the real world, object can always be characterized by multiple modalities. For example, in action recognition, one can integrate data from video, audio, and motion sensors to identify various human actions [1]. Similarly, in article classification, predictions can be made by comprehensively fusing both content and images [2]. Compared with single-modal data, multi-modal data is more informative and covers a wider range of information dimensions and diversity. Hence, it is more important to use multiple modal data to perceive the world. By leveraging multi-modal data, multi-modal learning strives to surpass single-modal learning, capturing widespread attention across diverse domains [3, 4, 5]. With the development of deep learning techniques [6, 7, 8], many multi-modal deep fusion networks have been proposed [9, 10, 11, 12]. They often employ different joint training strategies such as feature interaction [12, 13, 14] (i.e., two modalities interact from the input level) and prediction ensemble [15, 16] (i.e., two modalities interact from the prediction level), all while optimizing a unified learning objective.

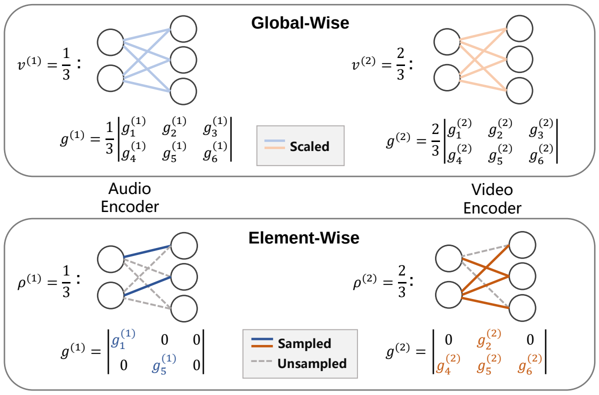

Global-Wise
$v^{(1)} = \frac{1}{3}$:
$v^{(2)} = \frac{2}{3}$:
$g^{(1)} = \frac{1}{3}\begin{vmatrix} g_1^{(1)} & g_2^{(1)} & g_3^{(1)} \\ g_4^{(1)} & g_5^{(1)} & g_6^{(1)} \end{vmatrix}$
Scaled
$g^{(2)} = \frac{2}{3}\begin{vmatrix} g_1^{(2)} & g_2^{(2)} & g_3^{(2)} \\ g_4^{(2)} & g_5^{(2)} & g_6^{(2)} \end{vmatrix}$
Audio Encoder
Video Encoder
Element-Wise
$\rho^{(1)} = \frac{1}{3}$:
$\rho^{(2)} = \frac{2}{3}$:
$g^{(1)} = \begin{vmatrix} g_1^{(1)} & 0 & 0 \\ 0 & g_5^{(1)} & 0 \end{vmatrix}$
Sampled
Unsampled
$g^{(2)} = \begin{vmatrix} 0 & g_2^{(2)} & 0 \\ g_4^{(2)} & g_5^{(2)} & g_6^{(2)} \end{vmatrix}$

Fig. 1: The illustration of different gradient modulation strategies using Audio-Video modalities as an example, where $v^{(1)}/v^{(2)}$ and $\rho^{(1)}/\rho^{(2)}$ denote the gradient modulation coefficient and the proportion of parameters to be updated, respectively, for the Audio/Video encoder. Under the two modulation strategies, $g^{(1)}/g^{(2)}$ represents the final updated gradient of the Audio/Video encoder. During backward propagation, **Global-wise** applies uniform modulation to the gradients (Scaled) for all parameters, while **Element-wise** involves differential modulation of the parameter gradients (Sampled or Unsampled).

Recent studies [17, 18] reveal that multi-modal approaches maybe perform far from their upper bound even though outperform the single-modal approaches, or even inferior to the single-modal model in certain situations [19]. This phenomenon is caused by the notorious “modality imbalance” problem [17, 20] during training, which involves the presence of a dominant modality and a non-dominant modality. Therefore, in the multi-modal joint training, due to the inherent greediness [21], the model updates excessively lean towards the dominant modalities, neglecting the learning of the non-dominant modality. Consequently, the non-dominant modality experiences severely slow learning, resulting in the performance of multi-modal learning inferior to that achieved in single-modal learning. Similar

- *Yang Yang, Hongpeng Pan, Qing-Yuan Jiang and Jinhui Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.*
  E-mail: {yyang, hongpengpancs, jiangqy, jinhuitang}@njust.edu.cn
- *Yi Xu is with the School of Control Science and Engineering, Dalian University of Technology, Dalian 116081, China.*
  E-mail: yxu@dlut.edu.cn
  Corresponding author: Yi Xu.

TABLE 1: Performance comparison of a multi-modal joint training model under global-wise and element-wise parameter update methods with different gradient modulation coefficients and proportions of updated parameters, respectively. The accuracy of the state-of-the-art method is 67.10. Results superior to the state-of-the-art methods are indicated by **bold**.

Global-wise		$v^{(2)}$				
(Accuracy)		1.0	0.8	0.6	0.4	0.2
	1.0	64.55	64.01	63.16	61.53	59.91
	0.8	64.74	64.28	63.00	62.27	60.53
$v^{(1)}$	0.6	65.01	64.74	63.97	62.62	60.38
	0.4	65.24	64.90	64.51	63.39	61.77
	0.2	66.13	65.28	65.13	63.89	63.20

Element-wise		$\rho^{(2)}$				
(Accuracy)		1.0	0.8	0.6	0.4	0.2
	1.0	64.55	63.74	63.32	62.93	59.72
	0.8	64.48	64.44	64.40	62.62	60.80
$\rho^{(1)}$	0.6	65.83	65.13	65.02	62.62	61.65
	0.4	**67.14**	66.33	66.60	63.86	62.85
	0.2	**68.96**	**68.53**	**68.84**	66.25	63.97

phenomena have been observed across various multi-modal tasks [19, 22, 23]. Therefore, the inefficiency in leveraging and fusing information from diverse modalities poses a significant challenge to the field of multi-modal learning.

To track this problem, initial work [19] found that different modalities may suffer from overfitting and converge at different rates, which leads to inconsistent learning efficiency when directly optimizing a uniform objective across different modalities. Furthermore, [21] introduced that the model tends to learn the dominant modality while neglecting the learning of non-dominant modalities, which affects the full utilization of multi-modal information. To cope with this issue, currently, the majority of studies [17, 18, 21, 24, 25] have been proposed to modulate each modal gradient during the back-propagation process, either assigning different learning rates to different modal branches or introducing additional losses for each modality. These strategies aim to maximize the contribution of each modality in multi-modal learning. Overall, these methods consistently utilize coarse-grained control at the modal level (global-wise) by updating the complete parameters of each modality. An example of global-wise updating mechanism is presented in the top panel of Figure 1, where $v^{(1)}/v^{(2)}$ represents the gradient modulation coefficient of the Audio/Video encoder. Within this mechanism, a uniform scaling factor is applied to scale gradients for parameters within the same modality, disregarding the distinctions in importance among different parameters. Inspired by recent advanced progress in subnetwork optimization [26, 27], we explore a fine-grained subnetworks optimization (element-wise updating) to update the gradients during backward procedure. In contrast to the global-wise updating mechanism, which utilizes the gradient modulation coefficient v to scale all gradients, the element-wise updating mechanism updates partial gradients according to a specified parameter update ratio ρ. As illustrated in the bottom panel of Figure 1, $\rho^{(1)}/\rho^{(2)}$ represents the proportion of parameters to be updated for the Audio/Video encoder, with the parameter selection strategy employing uniform sampling. We compare the global-wise and element-wise updating mechanisms through a preliminary experiment. Specifically, we adopt a multi-modal classification task and take the Kinetics-Sound dataset [28] which includes audio (dominant) and video (non-dominant) modalities as an example. The multi-modal network employs concatenation fusion at the last layer of each uni-modal stream before prediction. Table 1 illustrates the comparison between random strength gradient global modulation across different modalities and parameter mask element-wise modulation. The results reveal that in the majority of cases, the element-wise modulation strategy outperforms its counterparts. In certain instances, it even surpasses the performance of current state-of-the-art methods.

Drawing inspiration from importance sampling [29], we intend to refine the direct uniform parameters sampling strategy by dynamically adapting it based on the training data. Building upon this concept, we turn to optimizing each modal subnetwork, thereby fine-grained stimulating the non-dominant modality and alleviating the suppression from the dominant modality. To this end, we design the Adaptively Mask Subnetworks strategy considering modal Significance (AMSS). Specifically, we first introduce a simple yet effective mechanism to capture the batch-level significance of each modality, by calculating the mutual information rate with each modal prediction. Different from existing imbalanced multi-modal learning that directly weights the entire parameter gradients of each modality, we mask different sizes of promising parametric subnetworks, through NAS for each modality based on the modal significance. Therefore, the non-dominant modality masks a smaller subnetwork, while the dominant modality masks a larger subnetwork. After selection, we then perform partial gradient updates for each modality after masking the gradient of subnetwork parameters. Different from pruning operations, our mask strategy involves differentially updating various parameters within the model during back-propagation. Throughout the model forward process, we still use all parameters to calculate the loss. Furthermore, we present a theoretical validation of the reliability of the AMSS optimization strategy via convergence analysis. Nonetheless, this validation is contingent upon certain assumptions, leading to biased estimation. To mitigate this issue, we introduce an enhanced optimization strategy termed AMSS+, grounded in unbiased estimation principles, thereby circumventing constraints imposed by specific assumptions. In summary, the main contributions of this paper are summarized as follows:

- Based on the preliminary experiment, we propose a novel element-wise updating method called AMSS to solve the modality imbalance problem. AMSS can fine-grained stimulate the non-dominant modality and alleviate the suppression from the dominant modality. To the best of our knowledge, this is the first work that adopts element-wise updating mechanism in multi-modal learning.
- We engage in theoretical analysis to showcase the effec-

tiveness of subnetwork update strategies in imbalanced multi-modal learning. Additionally, drawing from theoretical findings, we introduce a novel mask strategy under unbiased estimation, termed AMSS+.

- We conduct extensive experiments across various modal scenarios, clearly demonstrating the effectiveness of fine-grained subnetworks optimization optimization in achieving a balanced learning approach for a multi-modal network.

2 Related Work

2.1 Multi-Modal Learning

Multi-modal learning aims to fuse different modalities from diverse sources [30, 31, 32]. Existing multi-modal methods commonly employ model-agnostic approaches, which can be classified based on fusion stages [30, 33] into early fusion [9, 12, 34, 35], late fusion [36, 37, 38, 39], and hybrid fusion [14, 40]. In detail, early fusion methods fuse features after extraction through either simple or meticulously designed networks. For example, concatenating features from different modalities to create a joint representation or applying affine transformations [9] to features for adaptive influence on the neural network's output has been explored. Additionally, late fusion, also termed prediction fusion, performs integration after each model makes predictions. Many studies have explored sophisticated late fusion methods beyond basic operations like summation and mean operations. For instance, [10] introduced an additional attention network for adaptive weight assignment to modality predictions, [16] employed an innovative Transformer-based late fusion architecture with bottlenecks as channels for inter-modal information interaction. Meanwhile, Hybrid fusion [14] endeavored to amalgamate the advantages of both approaches within the architecture and enhance the late fusion framework by incorporating a multi-modal transfer module to increase the interaction among features. However, the effectiveness of these approaches relies on the assumption that each modality makes a sufficient contribution throughout the joint training process.

2.2 Imbalanced Multi-Modal Learning

Recent studies [17, 18, 19, 25, 41, 42] have demonstrated that despite having access to more information in multi-modal learning, the performance improvement is often limited, and in some cases even worse than single-modal learning. Therefore, imbalanced multi-modal learning aims to rebalance the fitting speeds of non-dominant and dominant modalities, ensuring that each modality is fully utilized during the model training process. Based on this idea, [19] proposed a gradient blending technique that assigns different weights to branches based on the overfitting behavior of each modality, which aims to achieve optimal gradient blending. [17, 25] utilized a dynamic gradient modulation strategy, which reduces the learning rate of the dominant modality. [18] attempted to stimulate non-dominant modalities using prototype cross-entropy and employ prototypical entropy regularization to reduce dominant modality suppression. However, they always directly control the entire parameters update of each modality with unified weight. Furthermore, several attempts tried to employ extra networks to help multi-modal learning. For example, [41] introduced additional uni-modal branches and distilled the features from these branches into the multi-modal network, [21] introduced a hierarchical interaction module based on inter-modal information gain for modal representation learning, which can stabilize the differences between modalities during the training process. Nevertheless, the addition of extra modules introduces intricacies into the training procedure.

2.3 Subnetwork Optimization

Subnetwork optimization strategies have demonstrated their effectiveness in mitigating the challenge of overfitting in deep neural networks. Current research in this domain can be categorized into two distinct approaches. The first category involves network pruning strategies during the forward process, exemplified by techniques such as DropConnect [26], Gaussian Dropout [43], and Spatial Dropout [44]. These methodologies entailed the pruning of network nodes or connections, thereby reducing the scale of neural networks. The second category of methods [27, 45, 46] entailed retaining the complete set of network parameters for learning during the forward pass, and then specific neuron gradients are strategically masked to prevent their updates during the back-propagation. Nonetheless, it is worth noting that all subnetwork optimization techniques are devised with a focus on individual training models and are difficult to extend to the joint training of multi-modal models under scenarios characterized by data imbalance.

3 Rebalance Multi-Modal Network

In this section, we introduce the mask subnetworks (MS) for multi-modal learning in subsection 3.1. Then we employ MS for the multi-modal model to balance the learning speed across different modalities. Consequently, we propose the method named Adaptively Mask Subnetworks considering modal significance (AMSS). In subsection 3.2, we elaborate on the adaptive construction process of mask subnetworks of the multi-modal model. In subsection 3.3, we provide the theoretical convergence of AMSS and additionally introduce AMSS+, which is based on unbiased estimation.

3.1 Preliminary

For simplicity, we use boldface lowercase letters like $\mathbf{z}$ to denote the vectors, and the i-th element of $\mathbf{z}$ is denoted as z_i. Boldface uppercase letters like $\mathbf{Z}$ denote matrices and the element in the i-th row and j-th column of $\mathbf{Z}$ is denoted as Z_{ij}. We use numerical superscripts inside parentheses to denote specific modality, e.g., $\mathbf{z}^{(k)}$ denotes the k-th modality of $\mathbf{z}$. The notation $\mathcal{S}^{(k)}$ represents the parameter subnetwork set for the k-th modality. $\mathbb{E}(\cdot)$, $\mathbb{H}(\cdot)$ and $\mathbb{I}(\cdot)$ denote the exception, information entropy and mutual information, respectively. Furthermore, we use the symbol $\odot$ to denote the Hadamard product (i.e., element-wise product) of vectors/matrices.

Without any loss of generality, we first represent the training set as $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_N, \mathbf{y}_N)\}$, each example $\mathbf{x}_i$ is with K modalities, i.e., $\mathbf{x}_i = \{\mathbf{x}_i^{(k)}\}_{k=1}^{K}$,

$\mathbf{y}_i \in \{0,1\}^C$, C is the class number. The goal is to use this dataset $\mathcal{D}$ to train a model that can predict $\mathbf{y}_i$ accurately.

Most multi-modal deep neural networks [16, 18, 47] adopt multiple separate branches for the final prediction. These branches consist of multiple feature encoders, $\{\varphi^{(k)}(\mathbf{x}_i^{(k)})\}_{k=1}^{K}$, which aim to extract representations from the $\mathbf{x}_i$ data for each modality. Then, the multi-modal fusion operation can be denoted as $\varphi(\mathbf{x}_i) = [\varphi^{(1)}(\mathbf{x}_i^{(1)}); \varphi^{(2)}(\mathbf{x}_i^{(2)}); \cdots ; \varphi^{(K)}(\mathbf{x}_i^{(K)})]^\top$. Therefore, the final prediction in the multi-modal approach can be expressed as:

$$q(\mathbf{x}_i) = f(\varphi(\mathbf{x}_i)),$$

where f denotes the classifier. Finally, our objective is to train the f, φ to predict $\mathbf{y}$ based on $\mathbf{x}$, by minimizing the loss between the prediction and the ground truths:

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

Considering the imbalanced multi-modal learning, previous research, such as OGM [17] and AGM [25], usually concentrate on modulating model gradients by manipulating learning rates, specifically assigning a lower gradient coefficient to dominant modalities, to rebalance multi-modal learning, which has been verified to be a valid way.

In the rest of this section, we provide the details of the gradient updating with rebalanced strategy. For simplicity, we adopt the k-th modality for illustrating and omit the superscript "(k)". We represent the parameters at the t-th iteration as $\mathbf{w}(t)$. The parameter for a specific modality is updated by stochastic gradient descent (SGD):

$$\mathbf{w}(t+1) = \mathbf{w}(t) - v(t) \cdot \eta \nabla \mathcal{L}(\mathbf{w}(t)),$$

where $\mathbf{w}$ denotes the vectorized parameters of our model, $\nabla \mathcal{L}(\mathbf{w}(t))$ is the gradient of loss function $\mathcal{L}(\mathbf{w})$ at $\mathbf{w}(t)$ and $\eta > 0$ is the learning rate. $v(t)$ is gradient modulation coefficient at the t-th iteration. If a modality is dominant, $v(t) < 1$. Otherwise, $v(t) = 1$ and the equation is equivalent to the standard parameter update method. However, these methods multiply the weight to each model gradient equally (but not all parameters contribute equally to the optimization of training), which we argue is sub-optimal and will lead to excessive computation. Therefore, we put forward MS that determines a gradient update subnetwork $\mathcal{S}(t)$ during the t-th iteration:

$$m_j(t) = \begin{cases} 1, & \text{if } w_j(t) \in \mathcal{S}(t), \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathbf{m}(t)$ is the 0/1 mask vector with the same size of $\mathbf{w}$. The parameter updating can be reformulated as follows:

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \mathcal{L}(\mathbf{w}(t)) \odot \mathbf{m}(t). \tag{1}$$

3.2 AMSS

Based on preliminary experiments, we found that the MS strategy with uniform sampling performs excellently as shown in Table 1. Additionally, inspired by importance sampling, which involves selecting parameters based on input data, we propose Adaptively Mask Subnetworks considering Modal Significance (AMSS), as illustrated in Figure 2. The most crucial challenge of AMSS is how to adaptively obtain $\mathcal{S}^{(k)}(t)$ for each modality. This process primarily encompasses two problems: 1) how many parameters need to be selected for each modality and 2) what are the criteria for selecting parameters. Further details will be elaborated in the following content.

3.2.1 Parameter Quantity Mask via Modal Significance

In most multi-modal learning tasks, both modalities are assumed to be predictive of the target. Therefore, information theory can naturally state the modal significance, in particular, the mutual information [48] $\mathbb{I}(\mathbf{X}^{(k)}; \mathbf{Y})$ (between each modality and ground-truth) measure how much information is shared between $\mathbf{X}^{(k)}$ and $\mathbf{Y}$, which can be viewed as how much knowing $\mathbf{Y}$ reduces our uncertainty of $\mathbf{X}^{(k)}$.

Furthermore, considering the strong correlation between mutual information and the information content in each modality, we turn to employ the straightforward yet effective mutual information rate for evaluation, which has placed greater emphasis on the inherent changes within each modality during the training process. The equation can be represented as:

$$\hat{u}^{(k)} = \frac{\mathbb{I}(\mathbf{X}^{(k)}; \mathbf{Y})}{\mathbb{H}(\mathbf{X}^{(k)})},$$

where $\mathbf{X}^{(k)}$ denotes a batch of k-th modal examples, $\mathbf{Y}$ denotes the set of corresponding ground-truths. $\mathbb{H}(\mathbf{X}^{(k)})$ denotes the information entropy of of k-th modality. Intuitively, when the mutual information rate of a modality is higher, it signifies that this modality has a significant impact on the predictive task. This is because this modality can reduce uncertainty in the predictive task, thereby enhancing predictive capability. However, due to the inherent challenge of directly estimating mutual information in high-dimensional space as [49], conducting such an estimation is practically infeasible. To facilitate the practical utilization of mutual information, we employ variational bounds to approximate its true value following [50]:

$$\mathbb{I}(\mathbf{X}^{(k)}; \mathbf{Y}) \geq \mathbb{H}(\mathbf{Y}) + \mathbb{E}_{p(\mathbf{x}^{(k)}, \mathbf{y})}[\log q(\mathbf{y} \mid \mathbf{x}^{(k)})], \tag{2}$$

$$\mathbb{E}_{p(\mathbf{x}^{(k)}, \mathbf{y})}[\log q(\mathbf{y} \mid \mathbf{x}^{(k)})] = \frac{1}{B}\sum_{i=1}^{B} \log q\left(\mathbf{y}_i \mid \mathbf{x}_i^{(k)}\right),$$

$$\mathbb{H}(\mathbf{Y}) = -\sum_{c=1}^{C} p(\mathbf{y}_c) \log p(\mathbf{y}_c),$$

where, B denotes the batch size, $\mathbb{H}(\mathbf{Y})$ represents the information entropy of the ground-truth $\mathbf{Y}$ and $p(\mathbf{y}_c) = \frac{1}{B}\sum_{i=1}^{B} y_{i_c}$. The Barber-Agakov lower bound in Equation 2 or is tight, i.e., there is no gap between the bound and truth value, when $p(\mathbf{y}_i \mid \mathbf{x}_i^{(k)}) = q(\mathbf{y}_i \mid \mathbf{x}_i^{(k)})$. Therefore, it is necessary to train a classifier predicting $q(\mathbf{y}_i \mid \mathbf{x}_i^{(k)})$ to approximate $p(\mathbf{y}_i \mid \mathbf{x}_i^{(k)})$. For different fusion methods, the computation of $q(\mathbf{y}_i \mid \mathbf{x}_i^{(k)})$ varies. In late fusion, $q(\mathbf{y}_i \mid \mathbf{x}_i^{(k)}) = \mathrm{softmax}(f^{(k)}(\varphi^{(k)}(\mathbf{x}_i^{(k)}))_{\hat{y}_i}$, where $\hat{y}_i = \mathrm{argmax}(\mathbf{y}_i)$ and each modality has its own classifier $f^{(k)}$ for prediction. In early and hybrid fusion, we use zero-padding to represent features excluding the k-th modality, i.e., $\{\varphi^{(n)}(\mathbf{x}_i^{(n)}) = \mathbf{0}^{(n)}\}_{n \in \mathcal{N}/\{k\}}$, where

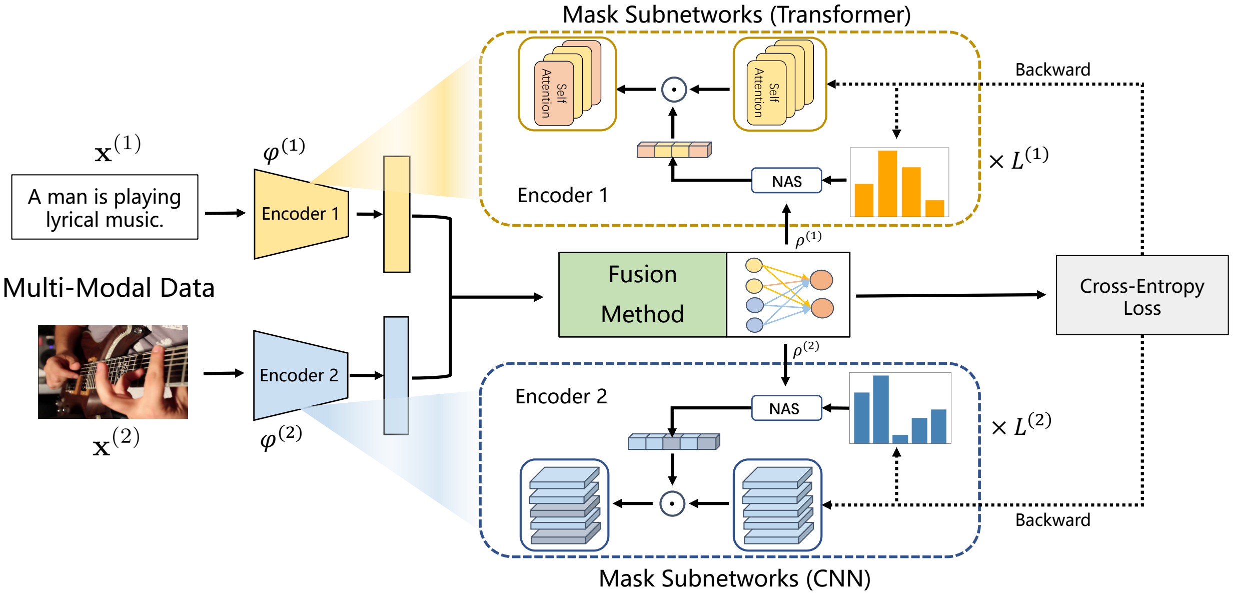


Fig. 2: The overall framework of our proposed AMSS strategy is illustrated using the Transformer-CNN structure as an example. $L^{(k)}$ denotes the number of layers in the k-th modality network. $\varphi^{(1)}$ and $\varphi^{(2)}$ represent the feature encoders for modality $\mathbf{x}^{(1)}$ and modality $\mathbf{x}^{(2)}$, respectively. $\rho^{(1)}/\rho^{(2)}$ denotes the proportion of parameters to be updated for the Audio/Video encoder, as determined through modal significance evaluation. NAS: non-uniform adaptive sampling.

$\mathcal{N} = \{1, 2, \ldots, K\}$ is the set of all modalities. Subsequently, $q(\mathbf{y}_i \mid \mathbf{x}_i^{(k)}) = \mathrm{softmax}(f(\varphi(\mathbf{x}_i)))_{\hat{y}_i}$.

Besides, to reduce fluctuations and stabilize the learning process, we employ the momentum update method, which evaluates the significance of the current modalities during the training phase by accumulating the modal significance obtained from the historical training data:

$$u^{(k)} = \lambda u^{(k)} + (1-\lambda)\hat{u}^{(k)}, \tag{3}$$

where λ is the attenuation factor. To address the problem of modality imbalance, we need to provide a greater advantage to the non-dominant modality during network updates while suppressing the dominant modality. Therefore, we mask fewer parameters for the non-dominant modality and comparatively more parameters for the dominant modality. Based on this idea, we first design the update ratio based on the modal significance $u^{(k)}$:

$$\rho^{(k)} = 1 - \frac{\exp(u^{(k)}/\tau)}{\sum_{n=1}^{K}\exp(u^{(n)}/\tau)}, \tag{4}$$

where τ serves as a hyper-parameter designed to adjust the size disparities across subnetworks of different modalities. Specifically, setting $\tau < 1$ amplifies these disparities, whereas $\tau > 1$ diminishes them. Furthermore, insights drawn from Table 1 suggest a preliminary conclusion: a higher ratio of non-dominant modal parameter updates relative to dominant ones often leads to improved model performance. Consequently, it is advisable to select $\tau < 1$ in experimental setups to amplify the disparities among subnetworks, a strategy that our subsequent experiments have confirmed to be effective.

3.2.2 Task-Guided Parameter Mask Criteria

As demonstrated in [45, 51, 52], it is evident that parameters with Fisher information estimation [53] play a pivotal role in learning the target task. Therefore, we adopt the Fisher information estimation as the selection criteria, which can provide an estimation of how much information a random variable carries about a parameter of the distribution [54], and measure the relative significance of parameters. In particular, the Fisher information of $\mathbf{w}^{(k)}$ is given by:

$$\mathbf{F}(\mathbf{w}^{(k)}) = \mathbb{E}\big[\big(\tfrac{\partial \log p(\hat{y}|\mathbf{x}^{(k)};\mathbf{w}^{(k)})}{\partial \mathbf{w}^{(k)}}\big)\big(\tfrac{\partial \log p(\hat{y}|\mathbf{x}^{(k)};\mathbf{w}^{(k)})}{\partial \mathbf{w}^{(k)}}\big)^{\top}\big].$$

Actually, $\mathbf{F}(\mathbf{w}^{(k)})$ can be viewed as the covariance of the gradient of the log-likelihood with respect to the parameters $\mathbf{w}^{(k)}$. Following [55], given the batch data, we use diagonal elements of the empirical Fisher information matrix to estimate the significance of parameters. Formally, we derive the Fisher information for the j-th parameter as follows:

$$F_j(\mathbf{w}^{(k)}) = \frac{1}{B}\sum_{i=1}^{B}\Big(\frac{\partial \log p(\hat{y}_i \mid \mathbf{x}_i^{(k)}; \mathbf{w}^{(k)})}{\partial w_j^{(k)}}\Big)^2.$$

Furthermore, we normalize the diagonal element as the importance of parameters:

$$p_j^{(k)} = \frac{F_j(\mathbf{w}^{(k)})}{\sum_{j=1}^{|\mathbf{w}^{(k)}|} F_j\left(\mathbf{w}^{(k)}\right)}.$$

As a result, the more important the parameter towards the target task, the higher $p_j^{(k)}$ it conveys. We set a probability distribution $\mathbf{P}^{(k)} = \{p_1^{(k)}, p_2^{(k)}, \ldots, p_{|\mathbf{w}^{(k)}|}^{(k)}\}$. Then, we employ the non-uniform adaptive sampling [56] without replacement for parameter selection. This approach allows $\mathcal{S}^{(k)}$ to concentrate on parameters associated with high information content. In contrast to the method of directly selecting the highest information parameters, it can encompass a more extensive scope of parameters through probabilistic sampling. Subsequently, based on the update ratio $\rho^{(k)}$, we conduct the sampling process to construct the parameter subnetwork set for each modality:

$$\mathcal{S}^{(k)} = \{s_1^{(k)}, s_2^{(k)}, \dots, s^{(k)}_{\lceil \rho^{(k)} * |\mathbf{w}^{(k)}| \rceil}\},$$

where $s_i^{(k)}$ represents a parameter in the k-th modal network that is selected during the i-th sampling. Subsequently, combining with Equation 1, we can derive parameter the gradient update strategy of AMSS as follows:

$$\mathbf{w}^{(k)}(t+1) = \mathbf{w}^{(k)}(t) - \eta \nabla \mathcal{L}(\mathbf{w}^{(k)}(t)) \odot \mathbf{m}^{(k)}(t).$$

To further improve the training efficiency for AMSS/AMSS+, we introduce the concept of the mask unit as a replacement for masking individual parameters and perform sampling based on this. Specifically, we design a channel-wise mask unit for CNN and a head-wise mask unit for Transformer. For other architectures, we continue to use an element-wise mask unit for sampling. The channel-wise mask unit for CNN is achieved by summing the importance scores of the remaining dimensions along the channel dimension to determine which channels should be masked. And a similar strategy is applied to the model along the head dimension for Transformer. Please refer to the appendix for more details.

3.3 Theoretical Analysis and Improved AMSS

In this subsection, we conduct a theoretical analysis of the convergence properties for updating parameters in the Mask Subnetwork under the non-convex optimization setting. Under the Mask-Incurred Error assumption, we have the following convergence result for AMSS. We include more details and the proof in Appendix.

Theorem 1 (Informal, AMSS). Under some assumptions for the stochastic gradient $\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 \le O\left(\frac{1+(1+\nu)^2}{\sqrt{T}(1+\nu)(1-\delta^2)}\right),$$

where $\delta \in (0,1)$ and $\nu \ge 0$ are two constants.

The result shows that AMSS converges to a stationary point with the rate of $O\left(\frac{1+(1+\nu)^2}{\sqrt{T}(1+\nu)(1-\delta^2)}\right)$, which is less worse than the $O\left(\frac{1+(1+\nu)^2}{\sqrt{T}(1+\nu)}\right)$ (i.e., $\delta = 0$, without masking) due to the Mask-Incurred Error assumption. To relax this assumption, we tend to propose another importance sampling strategy.

To this end, we propose a novel mask strategy that eliminates the bias of stochastic gradient $\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)$.

$$\hat{m}_j(t) = \begin{cases} \frac{1}{p_j}, & \text{if } w_j^{(k)}(t) \in \mathcal{S}^{(k)}(t), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Similarly, we have the following convergence result for the new smapling method.

Theorem 2 (Informal, AMSS+). Under some assumptions for the stochastic gradient $\nabla \ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 \le O\left(\frac{1+(1+\nu)^2}{\sqrt{T}(1+\nu)}\right),$$

where $\delta \in (0,1)$ and $\nu \ge 0$ are two constants.

Inspired by the theoretical foundation, we propose an improved method AMSS+. However, when the dimension of $\mathbf{m}$ is high, the value of p_j is usually too small so that $\frac{1}{p_j}$ is too large, which may lead to gradient explosion. Thus, in practice, we replace $\frac{1}{p_j}$ by $\frac{1}{p_j+a}$, where $a \in [0,1)$ is a constant used to eliminate gradient explosion. To facilitate the selection of the hyper-parameter a, we directly substitute the number of mask units $C_l^{(k)}$ of l-th layer in the network for a. Therefore, the Equation 5 can be modified to:

$$\hat{m}_j^{(k)}(t) = \begin{cases} \frac{1}{p_j^{(k)}+C_l^{(k)}}, & \text{if } w_j^{(k)}(t) \in \mathcal{S}^{(k)}(t), \\ 0, & \text{otherwise.} \end{cases}$$

Considering mask subnetworks under the unbiased estimation theory, we further propose the parameter gradient update strategy of AMSS+ as follows:

$$\mathbf{w}^{(k)}(t+1) = \mathbf{w}^{(k)}(t) - \eta \nabla \mathcal{L}(\mathbf{w}^{(k)}(t)) \odot \hat{\mathbf{m}}^{(k)}(t).$$

In summary, the aforementioned paragraph elucidates the strategies of both AMSS and AMSS+, with the latter distinguished by its utilization of the $0/\frac{1}{p_j}$ masking approach.

4 Experiment

4.1 Experimental Setups

Datasets. Following the prior research considering multimodal imbalance [17, 18], we adopt the **Kinetics-Sound** [28] and **CREAM-D** [57] datasets for validation, which includes audio and video modalities. To further validate the effectiveness of the proposed method, our research is extended in two dimensions. Firstly, the analysis is expanded to encompass the text-image modality, incorporating the **Sarcasm Detection** [58] and **Twitter-15** [59] datasets. Secondly, we employ the **NVGesture** [60] dataset to conduct research that goes beyond the limitation of two modalities.

Kinetics-Sound is used for video action recognition. It comprises 31 human action categories. The dataset contains a total of 19,000 10-second video clips (15k training set, 1.9k validation set, 1.9k test set). CREMA-D is designed for speech emotion recognition. It consists of 7,442 original clips. These clips are divided into 6,698 samples for the training set and 744 samples for the test set. CREMA-D can be categorized into six emotions. Sarcasm-Detection is designed for the task of sarcasm detection. It consists of 24,635 text-image pairs (19,816 training set, 2,410 validation set, 2,409 test set). The dataset is categorized into two classes. Twitter-15 is used for emotion recognition tasks, consisting of three classes. The data is collected from Twitter data [61], which comprises 5338 text-image pairs (3,179 training set, 1,122 validation sets, and 1,037 test set). NVGesture is collected using multiple sensors to investigate human-computer interfaces. It encompasses 1532 dynamic hand gestures (1,050 training set and 482 test set). We randomly sample 20% of training examples as the validation set, following [21], and use RGB, depth, and optical flow modalities in our setting. This dataset comprises 25 classes of hand gestures.

Baselines. We compare two categories of methods: 1) fusion methods considering modal rebalancing strategies, and 2) traditional fusion methods. In detail, modal rebalancing

TABLE 2: Comparison between AMSS with other SOTA methods on four datasets. The optimal performances are highlighted in **bold**. The underscore symbol represents the second best performance.

Methods		Kinetics-Sound		CREMA-D		Sarcasm-Detection		Twitter-15	
		ACC	mAP	ACC	mAP	ACC	Mac-F1	ACC	Mac-F1
Uni-modal	Audio/Text	54.12	56.69	63.17	68.61	81.36	80.65	73.67	68.49
	Video/Image	55.62	58.37	45.83	58.79	71.81	70.73	58.63	43.33
Multi-modal Fusion	Concat	64.55	71.30	63.31	68.41	82.86	82.40	70.11	63.86
	Affine	64.24	69.31	66.26	71.93	82.40	81.88	72.03	59.92
	Channel	63.51	68.66	66.13	71.70	-	-	-	-
	ML-LSTM	63.94	69.02	62.90	64.73	82.77	82.05	70.68	65.64
	Sum	64.90	71.03	63.44	69.08	82.94	82.47	73.00	66.61
	Weight	65.33	71.10	66.53	73.26	82.65	82.19	72.42	65.16
	ETMC	65.67	72.50	65.86	71.34	83.69	83.23	73.96	67.39
Multi-modal Rebalance Fusion	MSES	64.71	70.63	61.56	66.83	84.18	83.60	71.84	66.55
	OGR-GB	67.10	71.39	64.65	68.54	83.35	82.71	74.35	68.69
	OGM-GE	66.06	71.44	66.94	71.73	83.23	82.66	74.92	68.74
	Greedy	66.52	72.81	66.64	72.64	-	-	-	-
	DOMFN	66.25	72.44	67.34	73.72	83.56	82.62	74.45	68.57
	MSLR	65.91	71.96	65.46	71.38	84.23	83.69	72.52	64.39
	PMR	66.56	71.93	66.59	70.30	83.60	82.49	74.25	68.60
	AGM	66.02	72.52	67.07	73.58	84.02	83.44	74.83	69.11
	AMSS	68.96	74.89	67.61	73.97	84.14	83.69	**75.89**	**69.81**
	AMSS+	**72.25**	**79.13**	**70.30**	**76.14**	**84.35**	**83.77**	75.12	69.23

fusion methods include ORG-GB [19], MSES [62], OGM-GE [17], Greedy [21], DOMFN [63], MSLR [24], PMR [18], AGM [25]. The traditional fusion methods encompass feature concatenation fusion, affine transformation fusion [9], channel-wise fusion [14], multi-layers lstm fusion [34], prediction summation fusion, prediction weight fusion [10] and ETMC [47]. For convenience in the description, we abbreviate these methods as Concat, Affine, Channel, ML-LSTM, Sum, and Weight. Concat involves concatenating multiple modal features to obtain high-dimensional features, which are then used for downstream tasks. Affine applies a feature-wise affine transformation to intermediate neural network features. Channel recalibrates channel features of different CNN streams through the squeeze and multi-modal excitation steps. ML-LSTM method employs the LSTM structure to fully integrate various modalities. Sum predicts by calculating the probability mean of various modal prediction results. Weight assigns a weight to each individual modal branch using an attention mechanism. ETMC dynamically evaluates the trustworthiness of each modality across various samples, ensuring dependable integration.

Evaluation Metrics. We use accuracy (Acc) and mean Average Precision (mAP) for audio-video datasets following [17]. For the text-image dataset and NVGesture dataset, we utilize the Acc and Macro F1-score (Mac-F1) following [58, 59]. The Acc measures the proportion of concordance between predicted outcomes and true labels. The Mac-F1 computes the average of F1 scores for each category, while the mAP calculates the mean of average precision for each category.

Implementation Details. In all our experiments, we use raw data as input. Following [17, 18], for the Kinetics-Sound and CREMA-D datasets, we use ResNet18 [64] as the backbone for both modalities. In detail, for the video modality, we extract 10 frames from video clips and uniformly sample 3 frames as the input. The input channels are changed from 3 to 1, as demonstrated in [65]. For the audio modality, we convert the data into spectrograms with a size of 257×1004 for Kinetics-Sound and 257x299 for CREMA-D, using the librosa [66]. For the backbone of the text-image datasets, we employ ResNet50 and BERT [6] for the image and text modality, respectively. We crop the image data to a size of 224×224 and set the maximum sequence length for text data to 128. Besides, For the NVGesture dataset, we follow the data preparation steps outlined in [21] and employ the I3D [67] as uni-modal branches. For optimization and other technical details, please refer to the Appendix.

TABLE 3: The performance on NVGesture dataset. The involved modalities are RGB, OF, and Depth. Baseline means prediction sum fusion with no gradient modulation strategy. The best results are highlighted in **bold** and the underscore symbol represents the second best performance.

Methods	NVGesture scratch		NVGesture pretrain	
	ACC	Mac-F1	ACC	Mac-F1
RGB	68.88	69.05	78.22	78.33
OF	64.11	64.34	78.63	78.65
Depth	80.50	80.41	81.54	81.83
Baseline	78.63	78.91	82.57	82.68
MSES	79.46	79.48	81.12	81.47
ORG-GB	81.95	82.07	82.99	83.05
MSLR	81.54	81.38	82.37	82.39
AGM	80.71	81.18	82.78	82.84
AMSS	82.57	82.65	82.99	83.08
AMSS+	**84.64**	**84.94**	**83.20**	**83.25**

TABLE 4: The backbone of the network is transformer-based (MBT). Comparing with other imbalanced multi-modal learning methods. The best results are highlighted in **bold**. ↓ indicates a performance decrease compared to the baseline of the MBT model.

Methods	Kinetics-Sound		CREMA-D		Kinetics-Sound		CREMA-D	
	scratch		scratch		pretrain		pretrain	
	ACC	mAP	ACC	mAP	ACC	mAP	ACC	mAP
MBT	58.52	62.32	54.17	55.26	79.03	85.71	78.63	87.44
MSES	59.22	63.58	54.44	58.47	79.67	86.50	78.36 (↓)	87.41 (↓)
OGR-GB	59.18	62.27 (↓)	54.70	57.16	79.67	85.00 (↓)	79.03	87.74
OGM-GE	57.67 (↓)	62.24 (↓)	54.03 (↓)	54.94 (↓)	78.59 (↓)	85.78	78.23 (↓)	87.63
DOMFN	58.68	62.82	53.63 (↓)	54.70 (↓)	79.40	85.81	79.44	87.49
MSLR	59.72	63.71	54.30	58.63	79.47	86.53	78.76	87.95
PMR	58.06 (↓)	61.71 (↓)	53.36 (↓)	55.78	78.78 (↓)	85.23 (↓)	78.76	87.41 (↓)
AGM	**60.34**	63.61	54.84	55.15 (↓)	79.36	85.56 (↓)	79.30	87.97
AMSS	59.37	63.53	55.38	57.08	79.51	85.94	79.44	87.95
AMSS+	60.03	**64.28**	**56.18**	**59.51**	**80.09**	**86.25**	**79.57**	**88.10**

4.2 Comparison with Multi-Momal Learning Methods

To substantiate the advantages of AMSS and AMSS+, we conduct a comprehensive comparison, considering both modal rebalancing methods and traditional fusion approaches. Moreover, we identify limitations in the model architecture of Channel and Greedy constraints, rendering them unsuitable for the transformer structure. Additionally, to ensure experimental fairness, we evaluate AMSS, AMSS+, and all comparative methods using the same backbone, and use Concat for all multi-modal rebalance fusion methods to ensure a unified fusion strategy. We employ experiments on NVGesture from scratch (scratch) and with pre-training (pretrain). Taking into account the limitations posed by specific comparative methods in scenarios of two modalities, we undertake comparisons on the NVGesture dataset, encompassing MESE, ORG-GB, MSLR, and AGM methods.

The results for both audio-video datasets and text-image datasets are presented in Table 2, while results of the NVGesture dataset are depicted in Table 3. From the results, we draw the following observations: (1) On both the Twitter-15 and NVGesture datasets, we observe the phenomenon where the best uni-modal performance surpasses that of multi-modal joint learning. Besides, on other datasets, fusion methods without rebalancing exhibit limited improvement compared to the best uni-modal performance, especially on the CREMA-D and Sarcasm-Detection datasets. This limitation stems from the challenge of modality imbalance. (2) All modal rebalancing methods exhibit substantial enhancements compared to the traditional feature concatenation fusion. This observation not only highlights the influence of the imbalance phenomenon on performance but also substantiates the effectiveness of the modal rebalance strategy. (3) It is evident that AMSS/AMSS+ consistently achieves superior performance across all metrics compared to other comparison methods. We observe a significant improvement in the performance of AMSS+ on Kinetics-Sound/CREMA-D. After modulating, our method achieves a performance improvement of 5.15%/2.96% and 7.70%/6.99% in the Accuracy metric compared to the second-best approach and Concat. (4) Differing from modal rebalancing methods restricted to scenarios with only two modalities, such as OGM-GE and Greedy, our approach can address challenges in scenarios involving more than two modalities. In the evaluation of the NVGesture dataset, AMSS+ consistently achieves the best performance compared to other methods designed for multiple modalities. It is worth noting that, unlike other methods, the effectiveness of the AMSS+ in training from scratch is even better than pre-training. This observation confirms the robustness and effectiveness of our proposed method. (5) Compared to the biased estimation approach of AMSS, AMSS+ with unbiased estimation demonstrates superior performance in most scenarios, especially on audio-video datasets. This improvement indicates that AMSS+ leads to more accurate parameter estimation and superior performance. The consistency between experimental results and theoretical predictions bolsters the credibility of our unbiased estimation strategy.

Comparision in Complex Transformer-based Architecture. Currently, numerous works in multi-modal learning are built upon a unified multi-modal transformer architecture. To assess the effectiveness of the AMSS/AMSS+ method within this framework, we investigate alternative backbone networks. While handling audio-video datasets, in addition to utilizing CNN as the backbone, we introduce a fusion architecture based on Transformer, namely MBT [16]. This methodology comprises cross-modal interaction layers utilizing bottlenecks to integrate information between modalities. We employ two training approaches: training from scratch (scratch) and pre-training using ImageNet/Audioset for the ViT/AST models (pretrain) in MBT. Based on the results shown in Table 4, the following conclusions can be observed. (1) The effectiveness of modality imbalance methods on this architecture is limited compared to CNN architecture. In complex cross-modal interaction scenarios, certain modality imbalance methods prove ineffective. For example, OGM-GE and PMR, regardless of whether they are employed in a from-scratch training or pre-training setting, exhibit performance even worse than the benchmark results of MBT. (2) Whether employing a CNN architecture or a sophisticated multi-modal Transformer architecture, the AMSS+ strategy consistently maintains superior performance across almost all metrics. This demonstrates that our method possesses excellent adaptability. (3) Whether the model is pre-training or not does not impact the perfor-

TABLE 5: Various fusion methods combined with AMSS. † and ‡ indicates that AMSS abd AMSS+ has been applied, respectively.

Methods	Kinetics-Sound		CREMA-D		Sarcasm-Detection		Twitter-15	
	ACC	mAP	ACC	mAP	ACC	Mac-F1	ACC	Mac-F1
Affine	64.24	69.31	66.26	71.93	82.40	81.88	72.03	59.92
Affine†	65.02	71.60	66.94	71.38	83.31	82.70	72.81	65.42
Affine‡	**69.08**	**74.88**	**69.76**	**78.15**	**83.40**	**82.74**	**73.10**	**67.06**
Channel	63.51	68.66	66.13	71.70	-	-	-	-
Channel†	65.71	72.03	67.74	**76.91**	-	-	-	-
Channel‡	**68.69**	**75.42**	**69.89**	75.67	-	-	-	-
ML-LSTM	63.94	69.02	62.90	64.73	82.77	82.05	70.68	65.64
ML-LSTM †	67.80	73.24	65.05	71.59	83.44	82.80	73.67	67.77
ML-LSTM ‡	**70.70**	**77.00**	**67.47**	**73.50**	**83.89**	**83.12**	**74.45**	**69.41**
Sum	64.90	71.03	63.44	69.08	82.94	82.47	73.00	66.61
Sum†	66.52	72.98	66.80	73.14	**83.73**	83.07	**73.87**	**66.84**
Sum‡	**69.93**	**76.26**	**69.49**	**74.80**	83.69	**83.14**	73.38	66.55
Weight	65.33	71.10	66.53	73.26	82.65	82.19	72.42	65.16
Weight†	66.64	72.88	68.41	**77.22**	83.23	82.59	73.48	67.40
Weight‡	**68.46**	**74.62**	**71.10**	76.73	**83.98**	**83.42**	**74.35**	**69.52**

mance of our method. This flexibility allows our approach to be applied seamlessly across various scenarios. (4) AMSS+ continues to outperform AMSS, aligning once again with our theoretical expectations.

Exploration on Different Fusion Strategy. In the aforementioned context, we noted that traditional fusion methods are affected by the challenge of modality imbalance compared to multi-modal rebalancing strategies. The problem of modality imbalance constrains the performance of these methods. Therefore, we investigate the efficacy of integrating AMSS/AMSS+ with various fusion techniques to tackle the challenge of modality imbalance under different fusion strategies. The AMSS/AMSS+ strategy is applied to five fusion methods: Affine, Channel, ML-LSTM, Sum, and Weight, all of which have been introduced previously. It is noteworthy that Sum and Weight are prediction-level fusion methods, Channel is hybrid fusion method, while the others operate at the feature level. Fusion strategies at different levels correspond to different methods for the selection scope of subnetworks for each modal parameter. The relevant content is introduced in implementation details. As shown in Table 5, the combination of AMSS/AMSS+ with either feature-level or prediction-level fusion methods reveals a significant enhancement in their performance, underscoring the effectiveness of the AMSS strategy in augmenting their capabilities and mitigating the problem of modality imbalance across diverse fusion strategies. Compared to AMSS, the rebalancing strategy of AMSS+ demonstrates superior performance improvement, except for the Sum strategy on the text-image dataset.

4.3 Ablation Study

4.3.1 Sampling Mechanism Variant

In this section, our objective is to delve into the significance of multi-modal subnetworks derived from non-uniform adaptive sampling techniques. To achieve this, we employ the uniform sampling mechanism (Random) to conduct ablation studies for AMSS+. In the Random method, similar to that in preliminary experiments, each mask unit is sampled with an identical probability. The difference lies in how the update ratios ρ for various modalities depend on the dynamic update strategy of AMSS+. As shown in Table 6, the Random method selects a multi-modal subnetwork that exhibits superior performance compared to the Baseline, highlighting the effectiveness of updating each modality with distinct update ratios depending on AMSS+. Across both biased and unbiased estimation scenarios, the utilization of non-uniform adaptive sampling showcases significant superiority over uniform sampling in performance metrics. Based on these observations, we can confidently conclude that AMSS+ with non-uniform adaptive sampling is indeed more effective for downstream tasks.

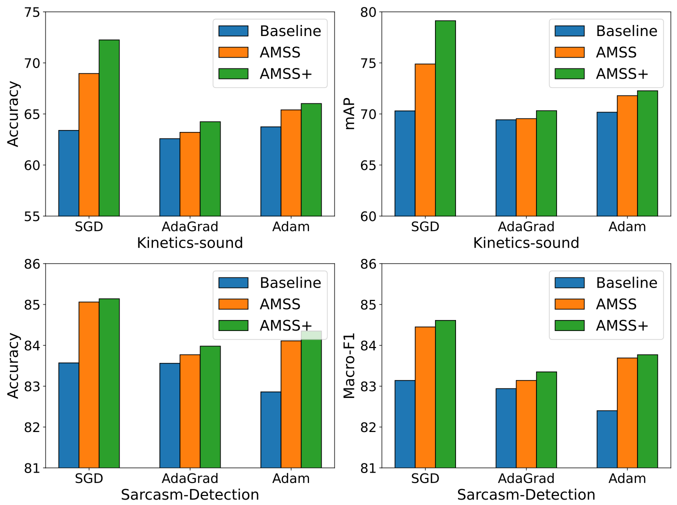


Fig. 3: Experiments with SGD, AdaGrad and Adam optimizers in Kinetics-Sound and Sarcasm-Detection. Baseline means no extra modulation.

4.3.2 Different Optimizer

Our theoretical analysis is grounded in the SGD optimizer. To further substantiate the adaptability of our approach when integrated with a variety of optimizers, AdaGrad [68]

TABLE 6: Performance of diverse sampling strategies. Baseline means no subnetwork optimization strategy. The best results are highlighted in **bold**.

Methods	Kinetics-Sound		CREMA-D		Sarcasm-Detection		Twitter-15	
	ACC	mAP	ACC	mAP	ACC	Mac-F1	ACC	Mac-F1
Baseline	64.55	71.30	63.31	68.41	82.86	82.40	70.11	63.86
Random	66.45	72.43	65.59	72.05	83.60	83.00	74.35	68.87
AMSS	68.96	74.89	67.61	73.97	84.14	83.69	**75.89**	**69.81**
AMSS+	**72.25**	**79.13**	**70.16**	**76.14**	**84.35**	**83.77**	75.12	69.23

TABLE 7: The effectiveness of each component.

Methods	Kinetics-Sound		Sarcasm-Detection	
	ACC	mAP	ACC	Mac-F1
Baseline	64.55	71.30	82.86	82.40
w/ Classifier Mask	65.37	70.87	83.40	83.05
w/ Backbone Mask	66.68	72.23	83.81	83.17
AMSS	**68.96**	**74.89**	**84.14**	**83.69**
w/ Classifier Mask+	66.80	73.80	83.44	82.93
w/ Backbone Mask+	69.15	76.13	83.77	83.10
AMSS+	**72.25**	**79.13**	**84.35**	**83.77**

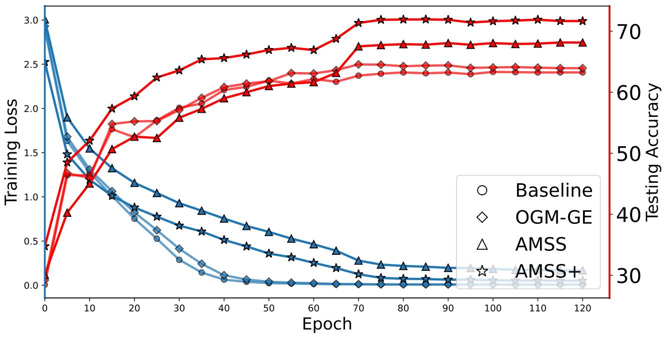


Fig. 4: On the Kinetics-Sound dataset, we employ the concatenation fusion method for the joint training of multi-modal models, encompassing Baseline, OGM-GE, AMSS, and AMSS+. We investigate the changes in training loss and evaluate the variations in test performance across these multi-modal models. Baseline means no gradient modulation strategy.

and Adam [69] optimizers are also employed in our experimental validation. These optimizers are applied to two distinct types of datasets to ensure a comprehensive evaluation of our approach's performance. Subsequently, we integrate the AMSS or AMSS+ strategy and assess its performance across different optimizers. By incorporating AMSS or AMSS+, we aim to enhance the performance of our method across a spectrum of optimization strategies. The results, depicted in Figure 3, emphasize the diverse performance exhibited by the selected optimizers on Kinetics-Sound and Sarcasm-Detection datasets. Significantly, our method consistently showcases exceptional adaptability, consistently surpassing baseline results and achieving substantial performance improvements. Furthermore, compared to other optimizers, the improvement in the effectiveness of our method is more pronounced under the SGD optimizer. This is attributed to the adaptive adjustment of learning parameters inherent to the optimizer itself. This sustained success across different optimizers underscores the adaptability of our approach in optimizing model performance, irrespective of the underlying optimizer.

4.3.3 Component Ablation Analysis

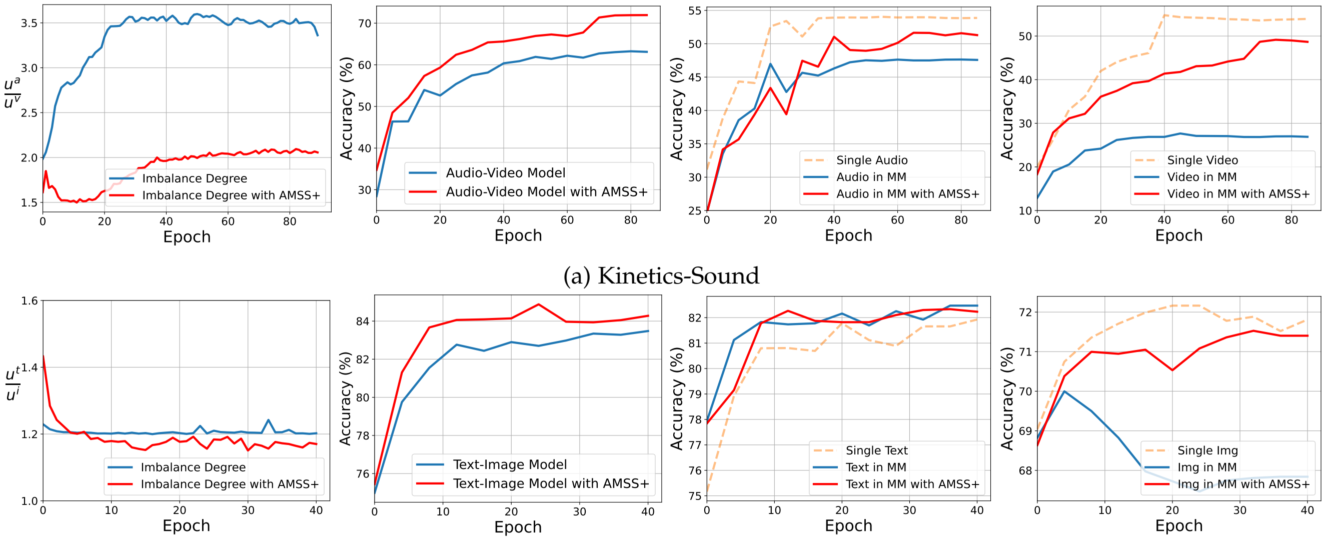
In both fusion methods at the prediction level and the feature concatenation of a single fully connected layer (considered decoupled) [17], we implement the adaptive subnetwork masking strategy independently for the classifier and the backbone network. To assess the significance of each module within the model, we perform an extensive module ablation analysis on two different types of datasets. Table 7 illustrates the impact of each module on our method. We present four model variants for examination: Baselines; Model with classifier masks/masks+, excluding backbone masks/masks+; Model with backbone masks/masks+, excluding classifier masks/masks+; The proposed AMSS/AMSS+. In our analysis, each module integrated into our method consistently demonstrates a contribution to the overall performance enhancement of the model. Compared to applying the masking strategy to the classifier, adopting the masking strategy for the backbone network leads to a more pronounced improvement in the model's performance.

4.4 Optimization Analysis

In our research, we scrutinize the impact of diverse parameter update strategies on model training, concentrating on a comparative study of four distinct methods employed for multi-modal joint learning in the Kinetics-Sound dataset. Figure 4 demonstrates that the implementation of Mask Subnetwork (AMSS, AMSS+) strategies appears to moderate the pace at which the model learns, in contrast to the baseline model. From the theoretical perspective, subnetwork optimization strategies are expected to moderately slow the model's rate of convergence, a trend that is consistent with the trajectory of training loss depicted in Figure 4. It is noteworthy that despite this deceleration, such strategies maintain robust performance levels and demonstrate superior generalization abilities. In contrast, while other methods may exhibit a faster modal convergence rate, they evidently succumb to model overfitting issues, lacking generalization abilities. Additionally, the AMSS+ strategy, which provides an unbiased approach to subnetwork optimization, surpasses its biased counterpart, showing improved convergence rates and overall performance. In conclusion, the adoption of adaptive strategies for updating subnetworks offers tangible benefits.

4.5 Analysis of Modality Imbalance

The issue of modality imbalance highlights a challenge in the process of multi-modal learning, where the dominance of one modality inhibits the exploration of distinctive features in other modalities. To further scrutinize the impact of our approach on model optimization, we initially define



(a) Kinetics-Sound

(b) Sarcasm-Detection

Fig. 5: Analysis of Modality Imbalance Problem. Each dataset is represented in Figures from left to right, depicting the variation in model imbalance degree, the comparison between our method and the Baseline model, the performance of single-modal branches in multi-modal trained models, and the performance of single-modal branches with AMSS+, including Audio/Text and Video/Img modalities. The fusion method used in the multi-modal model is Concat.

TABLE 8: Accuracy and gradient update ratios across two modalities under varying τ.

AMSS+	$\tau < 1$										$\tau \geq 1$		
	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	2.00	4.00
ACC	72.17	**72.52**	72.25	72.13	71.82	71.47	71.67	71.05	71.01	70.85	70.20	69.85	69.69
$\overline{\rho}^{(1)}$	0.2238	0.2436	0.2479	0.2508	0.2684	0.2774	0.2914	0.2940	0.3105	0.3156	0.3305	0.3964	0.4404
$\overline{\rho}^{(2)}$	0.7762	0.7564	0.7521	0.7492	0.7316	0.7226	0.7086	0.7060	0.6895	0.6844	0.6695	0.6036	0.5596

the degree of modality imbalance, denoted as $\frac{u^{(1)}}{u^{(2)}}$, utilizing the modal significance derived from Equation 3. We use $\frac{u^a}{u^v}$ and $\frac{u^t}{u^i}$ to represent the imbalance degree of the audio-video and text-image modalities, respectively. Subsequently, we analyze the performance of the multi-modal model both before and after the application of the AMSS+ strategy, along with an examination of the performance of uni-modal branches within the multi-modal model. The observations and conclusions from the imbalance analysis are presented in Figure 5. Due to space limitations, the results of CREMA-D and Twitter-15 datasets are presented in the Appendix. (1) In vanilla multi-modal learning, the Audio/Text branch closely approximates the performance of the single Audio/Text model, while the Video/Img branch exhibits minimal effective learning, resulting in a noticeable gap compared to the single Video/Img modality. As depicted in the curve of imbalance degree variations, our method effectively alleviates the modality imbalance to varying degrees across the two datasets. This serves as concrete evidence of the efficacy of our approach. (2) The modality imbalance issue is most severe in the Kinetics-Sound, leading to under-utilization of the video modality in joint training. The performance of the multi-modal model is primarily derived from the Audio modality branch. The AMSS+ strategy allows for the full utilization of the Video modality, bringing its performance close to single-modal training. (3) On the text-img dataset, where the modality imbalance issue is less severe than in audio-video, the model is still able to address this problem, leading to improvements in the performance of non-dominant modalities (image) and ultimately achieving better overall model performance.

4.6 Exploring the Gradient Update Ratio

AMSS/AMSS+ introduces one hyper-parameter τ in Equation 4. In this section, we vary τ within the set $\{0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 4\}$ on Kinetics-sound to explore its impact on model performance and the variations in gradient update ratios across two modalities. Since the gradient update ratio in our method is on the batch level, we calculate the average gradient update ratio for each modality across all iterations within the initial 10 epochs, denoted here as $\overline{\rho}$. $\overline{\rho}^{(1)}/\overline{\rho}^{(2)}$ represent the Audio/Video modalities. As illustrated in Table 8, we observe that even when $\tau = 1$, indicating the absence of τ, the performance of AMSS+ still surpasses the results of the state-of-the-art method (67.10). Moreover, when $\tau < 1$, there is a pronounced size difference in the gradient update ratio among modalities, facilitating a better balance in optimization across modalities, thereby enhancing the overall efficacy of the model. Conversely, when $\tau > 1$, this size difference is markedly reduced, and the issue of modality imbalance remains significant, thus limiting the performance of the model. Intriguingly, it is observed that the size difference between modalities should not be excessively large as it could detrimentally impact the model's performance, potentially by inhibiting the update momentum of the dominant modality, as exemplified when $\tau = 0.1$. Moreover, through meticulous hyper-parameter modulation, AMSS+ demonstrates an ability to surpass previous benchmarks, achieving enhanced outcomes. For example, at $\tau = 0.2$, AMSS+ exhibits a performance increment of 0.27 compared to our previous results in the Kinetics-sound dataset. Additionally, when fixing $\overline{\rho}^{(1)} = 1$ and $\overline{\rho}^{(2)} = 1$, i.e., not masking the parameters, and scaling the parameters only with $\frac{1}{p_j + L_i}$ in AMSS+, the model performance is 69.39. This indicates that the masking subnetwork strategy in AMSS+ is crucial.

5 Conclusion

In our study, we found that the strategy of employing uniformly weighted modulation parameters for gradient modifications is sub-optimal for resolving the challenges posed by modality imbalance. To address this issue more effectively, we introduce the innovative approach of Adaptively Mask Subnetworks considering Modal Significance (AMSS). Our approach emphasizes performing gradient updates on differently sized, more promising subnetworks, selected adaptively for each modality, based on the batch-level mutual information rate. We conduct a theoretical analysis of effectiveness of this strategy and further propose an unbiased estimation optimization strategy, AMSS+. Furthermore, our approach can serve as a flexible plug-in strategy for existing multi-modal models.

Acknowledgments

This work was supported in part by the National Key RD Program of China (2022YFF0712100), in part by the NSFC (62276131), in part by the Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), in part by the Fundamental Research Funds for the Central Universities (No.30922010317), and in part by the Major Science and Technology Projects in Jiangsu Province under Grant (BG2024042).

References

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *TPAMI*, no. 01, pp. 1–20, 2022.

[2] J. C. Pereira and N. Vasconcelos, "Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems," *CVIU*, vol. 124, pp. 123–135, 2014.

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

[4] Q. Lu, Y. Long, X. Sun, J. Feng, and H. Zhang, "Fact-sentiment incongruity combination network for multimodal sarcasm detection," *INFORM FUSION*, vol. 104, p. 102203, 2024.

[5] R. Yan, L. Xie, X. Shu, L. Zhang, and J. Tang, "Progressive instance-aware feature learning for compositional action recognition," *TPAMI*, vol. 45, no. 8, pp. 10 317–10 330, 2023.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[7] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding." *TPAMI*, vol. 41, pp. 2070–2083, 2019.

[8] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *TPAMI*, vol. 44, pp. 3300–3315, 2022.

[9] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018, pp. 3942–3951.

[10] Y. Yang, K.-T. Wang, D.-C. Zhan, H. Xiong, and Y. Jiang, "Comprehensive semi-supervised multi-modal learning." in *IJCAI*, 2019, pp. 4092–4098.

[11] Y. Yang, J. Yang, R. Bao, D. Zhan, H. Zhu, X. Gao, H. Xiong, and J. Yang, "Corporate relative valuation using heterogeneous multi-modal graph neural network," *TKDE*, vol. 35, no. 1, pp. 211–224, 2023.

[12] X. Liang, Y. Qian, Q. Guo, H. Cheng, and J. Liang, "AF: an association-based fusion method for multi-modal classification," *TPAMI*, vol. 44, no. 12, pp. 9236–9254, 2022.

[13] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *TMM*, vol. 25, pp. 3375–3385, 2023.

[14] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: multimodal transfer module for CNN fusion," in *CVPR*, 2020, pp. 13 289–13 299.

[15] H. Ma, Z. Han, C. Zhang, H. Fu, J. T. Zhou, and Q. Hu, "Trustworthy multimodal regression with mixture of normal-inverse gamma distributions," in *NIPS*, 2021, pp. 6881–6893.

[16] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *NIPS*, 2021, pp. 14 200–14 213.

[17] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8238–8247.

[18] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in *CVPR*, 2023, pp. 20 029–20 038.

[19] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*, 2020, pp. 12 695–12 705.

[20] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably)," in *ICML*, 2022, pp. 9226–9259.

[21] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *ICML*, 2022, pp. 24 043–24 055.

[22] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: A multilayer approach for multimodal fusion," in *ECCV*, vol. 11134, 2018, pp. 575–589.

[23] J. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," *CoRR*, vol. abs/2006.15955, 2020.

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

[25] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *ICCV*, 2023, pp. 22 214–22 224.

[26] L. Wan, M. D. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, "Regularization of neural networks using dropconnect," in *ICML*, 2013, pp. 1058–1066.

[27] C. Lee, K. Cho, and W. Kang, "Mixout: Effective regularization to finetune large-scale pretrained language models," in *ICLR*, 2020.

[28] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017, pp. 609–617.

[29] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *international conference on machine learning*. PMLR, 2015, pp. 1–9.

[30] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.

[31] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *ISPREG*, vol. 34, no. 6, pp. 96–108, 2017.

[32] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *JSTSP*, vol. 14, no. 3, pp. 478–493, 2020.

[33] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *MULTIMEDIA SYST*, vol. 16, no. 6, pp. 345–379, 2010.

[34] W. Nie, Y. Yan, D. Song, and K. Wang, "Multimodal feature fusion based on multi-layers LSTM for video emotion recognition," *MULTIMED TOOLS APPL*, vol. 80, no. 11, pp. 16 205–16 214, 2021.

[35] Y. Zeng, W. Yan, S. Mai, and H. Hu, "Disentanglement translation network for multimodal sentiment analysis," *INFORM FUSION*, vol. 102, p. 102031, 2024.

[36] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao, "Late fusion incomplete multiview clustering," *TPAMI*, vol. 41, no. 10, pp. 2410–2423, 2018.

[37] Y. Du, Y. Wang, J. Hu, X. Li, and X. Chen, "An emotion role mining approach based on multiview ensemble learning in social networks," *INFORM FUSION*, vol. 88, pp. 100–114, 2022.

[38] M. Alfaro-Contreras, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Late multimodal fusion for image and audio music transcription," *EXPERT SYST APPL*, vol. 216, p. 119491, 2023.

[39] B. Liu, L. He, Y. Xie, Y. Xiang, L. Zhu, and W. Ding, "Minjot: Multimodal infusion joint training for noise learning in text and multimodal classification problems," *INFORM FUSION*, vol. 102, p. 102071, 2024.

[40] X. Zheng, C. Tang, Z. Wan, C. Hu, and W. Zhang, "Multi-level confidence learning for trustworthy multimodal classification," in *AAAI*, 2023, pp. 11 381–11 389.

[41] C. Du, T. Li, Y. Liu, Z. Wen, T. Hua, Y. Wang, and H. Zhao, "Improving multi-modal learning with unimodal teachers," *CoRR*, vol. abs/2106.11059, 2021.

[42] Y. Sun, S. Mai, and H. Hu, "Learning to balance the learning rates between various modalities via adaptive tracking factor," *SPL*, vol. 28, pp. 1650–1654, 2021.

[43] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *NIPS*, 2015, pp. 2575–2583.

[44] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *CVPR*, 2015, pp. 648–656.

[45] R. Xu, F. Luo, Z. Zhang, C. Tan, B. Chang, S. Huang, and F. Huang, "Raise a child in large language model: Towards effective and generalizable fine-tuning," in *EMNLP*, 2021, pp. 9514–9528.

[46] H. Zhang, G. Li, J. Li, Z. Zhang, Y. Zhu, and Z. Jin, "Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively," in *NIPS*, 2022, pp. 21 442–21 454.

[47] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multiview classification with dynamic evidential fusion," *TPAMI*, vol. 45, no. 2, pp. 2551–2566, 2023.

[48] K. Sridharan and S. M. Kakade, "An information theoretic framework for multi-view learning," in *COLT*, 2008, pp. 403–414.

[49] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, R. D. Hjelm, and A. C. Courville, "Mutual information neural estimation," in *ICML*, 2018, pp. 530–539.

[50] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *ICML*, 2020, pp. 1779–1788.

[51] J. Sourati, M. Akçakaya, D. Erdogmus, T. K. Leen, and J. G. Dy, "A probabilistic active learning algorithm based on fisher information ratio," *TPAMI*, vol. 40, no. 8, pp. 2023–2029, 2018.

[52] S. P. Singh and D. Alistarh, "Woodfisher: Efficient second-order approximation for neural network compression," in *NIPS*, 2020.

[53] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Phil. Trans.*, vol. 222, no. 594-604, pp. 309–368, 1922.

[54] M. Tu, V. Berisha, Y. Cao, and J. Seo, "Reducing the model order of deep neural networks using information theory," in *ISVLSI*, 2016, pp. 93–98.

[55] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *PNAS*, vol. 114, no. 13, pp. 3521–3526, 2017.

[56] S. Gopal, "Adaptive sampling for sgd by exploiting side information," in *ICML*, 2016, pp. 364–372.

[57] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: crowdsourced emotional multimodal actors dataset," *IEEE T AFFECT COMPUT*, vol. 5, no. 4, pp. 377–390, 2014.

[58] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *ACL*, 2019, pp. 2506–2515.

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

[60] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *CVPR*, 2016, pp. 4207–4215.

[61] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive

co-attention network for named entity recognition in tweets," in *AAAI*. AAAI Press, 2018, pp. 5674–5681.

[62] N. Fujimori, R. Endo, Y. Kawai, and T. Mochizuki, "Modality-specific learning rate control for multimodal classification," in *ACPR*, 2020, pp. 412–422.

[63] Y. Yang, J. Zhang, F. Gao, X. Gao, and H. Zhu, "DOMFN: A divergence-orientated multi-modal fusion network for resume assessment," in *ACM MM*, 2022, pp. 1612–1620.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[65] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020, pp. 721–725.

[66] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *SciPy*, 2015, pp. 18–24.

[67] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, 2017, pp. 4724–4733.

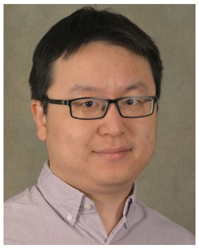
[68] J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J MACH LEARN RES*, vol. 12, no. 7, pp. 2121–2159, 2011.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

Yang Yang received the Ph.D. degree in computer science, Nanjing University, China in 2019. At the same year, he became a faculty member at Nanjing University of Science and Technology, China. He is currently a Professor with the school of Computer Science and Engineering. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining. He has published prolifically in refereed journals and conference proceedings, including IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Information Systems (ACM TOIS), ACM Transactions on Knowledge Discovery from Data (TKDD), ACM SIGKDD, ACM SIGIR, WWW, IJCAI, and AAAI. He was the recipient of the the Best Paper Award of ACML-2017. He serves as PC in leading conferences such as IJCAI, AAAI, ICML, NeurIPS, etc.

Hongpeng Pan is currently working towards the M.S. degree at the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly lie in deep learning and data mining. He is currently focusing on multi-modal learning.

Qing-Yuan Jiang received the BSc and the PhD degrees in computer science from Nanjing University, China. He has published 9 papers in leading international journal/conferences. He serves as a PC member in leading conferences such as AAAI, IJCAI, etc. He is currently an associate professor at Nanjing University of Science and Technology. His research interests are in learning to hash and multi-modal retrieval.



Yi Xu received the PhD degree in computer science from the University of Iowa, Iowa City, Iowa, USA. He is currently a professor with the School of Control Science and Engineering, Dalian University of Technology, China. His research interests are machine learning, optimization, deep learning, and statistical learning theory. He has published more than twenty papers in refereed journals and conference proceedings, including IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Transactions on Machine Learning Research, NeurIPS, ICML, ICLR, CVPR, AAAI, IJCAI, UAI. He serves as a SPC/PC/Reviewer in leading conferences such as NeurIPS, ICML, ICLR, CVPR, AAAI, IJCAI, etc.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. He has authored more than 200 articles in top-tier journals and conferences. His research interests include multimedia analysis and computer vision. Dr. Tang was a recipient of the Best Paper Awards in ACM MM 2007, PCM 2011, ICIMCS 2011, and ACM MM Asia 2020, the Best Paper Runner-Up in ACM MM 2015, and the Best Student Paper Awards in MMM 2016 and ICIMCS 2017. He has served as an Associate Editor for the IEEE TMM, IEEE TNNLS, the IEEE TKDE, and the IEEE TCSVT. He is a Fellow of IAPR.

Appendix A Convergence Analysis

In this subsection, we provide a convergence analysis for the proposed method under the non-convex optimization setting. The detailed process of the proof is as follows.

A.1 Case I: Biased Stochastic Gradient

For the first case of biased stochastic gradient, recall that the update step of stochastic gradient descent (SGD) is

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t) \tag{6}$$

where $\nabla \ell(\mathbf{w}(t))$ is the stochastic version of the gradient of loss function $\nabla \mathcal{L}(\mathbf{w}(t))$ at θ_t and $\eta > 0$ is the learning rate, and the element of $\mathbf{m}(t)$ is given by

$$m_j(t) = \begin{cases} 1, & \text{if } w_j(t) \in \mathcal{S}(t), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

To analyze the convergence rate, we make the following common assumptions for the loss function.

Assumption 1 (Smoothness). We assume that the loss function $\mathcal{L}$ is L-smooth. That is, for any $\mathbf{w}, \mathbf{w}'$, we have

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}') \leq \langle \nabla \mathcal{L}(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{L}{2} \|\mathbf{w} - \mathbf{w}'\|^2. \tag{8}$$

Based on Equation 1, we suppose that the stochastic gradient $\nabla \ell(\mathbf{w}(t))$ is unbiased, i.e., $\mathbb{E}[\nabla \ell(\mathbf{w}(t))] = \nabla \mathcal{L}(\mathbf{w}(t))$, which is commonly used in non-convex optimization. However, under the $0/1$ mask strategy, $\nabla \ell(\mathbf{w}(t))$ and $\mathbf{m}(t)$ are not independent, the stochastic gradient $\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)$ is biased, that is, $\mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)] \neq \nabla \mathcal{L}(\mathbf{w}(t))$. We make the following assumption for the stochastic gradient $\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)$.

Assumption 2 (Bounded Variance). We assume that the stochastic gradient $\nabla \ell(\mathbf{w}) \odot \mathbf{m}(t)$ is biased and its variance is bounded. That is, for any $\mathbf{w}(t)$ and $\mathbf{m}(t)$, we have

$$\mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)] = \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)), \tag{9}$$

and

$$\mathbb{E}[\|\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t) - \mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)]\|]^2 \leq \nu \|\nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\|^2 + \sigma^2, \tag{10}$$

where $\sigma^2 \geq 0$ and $\nu \geq 0$ are two constants.

Assumption 3 (Mask-Incurred Error). For any $\mathbf{w}(t)$ and $\mathbf{m}(t)$, we have

$$\|\mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t) - \nabla \ell(\mathbf{w}(t))]\| \leq \delta \|\mathbb{E}[\nabla \ell(\mathbf{w}(t))]\|, \tag{11}$$

where the constant $\delta \in [0, 1]$.

We present the convergence property in the following theorem.

Theorem 3 (Formal, AMSS). Under Assumptions 1, 2, 3, if the learning rate is set as $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$, then

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2(1+\nu)L\mathcal{L}(\mathbf{w}(1))}{\sqrt{T}(1-\delta^2)} + \frac{\sigma^2}{(1+\nu)\sqrt{T}(1-\delta^2)}. \tag{12}$$

Proof 1. By Assumption 1, we have

$$\begin{aligned} \mathcal{L}(\mathbf{w}(t+1)) - \mathcal{L}(\mathbf{w}(t)) \leq & \langle \nabla \mathcal{L}(\mathbf{w}(t)), \mathbf{w}(t+1) - \mathbf{w}(t) \rangle + \frac{L}{2} \|\mathbf{w}(t+1) - \mathbf{w}(t)\|^2 \\ \overset{(6)}{=} & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \mathbf{g}(t) \rangle + \frac{\eta^2 L}{2} \|\mathbf{g}(t)\|^2 \end{aligned} \tag{13}$$

where $\mathbf{g}(t) := \nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)$. Taking expectation over both sides of (13) and by using Assumption 2, we have

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\mathbf{w}(t+1)) - \mathcal{L}(\mathbf{w}(t))] \leq & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta^2 L}{2} \mathbb{E}[\|\mathbf{g}(t)\|^2] \\ = & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta^2 L}{2} (\mathbb{E}[\|\mathbf{g}(t) - \mathbb{E}[\mathbf{g}(t)]\|^2] + \mathbb{E}[\|\mathbb{E}[\mathbf{g}(t)]\|^2]) \\ \leq & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta^2 L}{2} \left((1+\nu)\|\nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\|^2 + \sigma^2\right) \\ \leq & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta}{2} \|\nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\|^2 + \frac{\eta^2 L \sigma^2}{2}, \end{aligned} \tag{14}$$

where the last inequality is due to $\eta \le \frac{1}{(1+\nu)L}$. Since $-\langle a, b\rangle + \frac{\|b\|^2}{2} = \frac{\|a-b\|^2}{2} - \frac{\|a\|^2}{2}$, then (14) implies that

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\mathbf{w}(t+1)) - \mathcal{L}(\mathbf{w}(t))] \le & -\eta\langle\nabla\mathcal{L}(\mathbf{w}(t)), \nabla\mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\rangle + \frac{\eta^2 L}{2}\mathbb{E}[\|\mathbf{g}(t)\|^2] \\ \le & \frac{\eta}{2}\|b(\mathbf{w}(t))\|^2 - \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2},\end{aligned} \tag{15}$$

Next, by (9) in Assumption 2 and (11) in Assumption 3, we know

$$\|b(\mathbf{w}(t))\| = \|\mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \mathbf{m}(t) - \nabla\ell(\mathbf{w}(t))]\| \le \delta\|\mathbb{E}[\nabla\ell(\mathbf{w}(t))]\| = \delta\|\nabla\mathcal{L}(\mathbf{w}(t))\|. \tag{16}$$

Therefore, by (15) and (16) we have

$$\mathbb{E}[\mathcal{L}(\mathbf{w}(t+1)) - \mathcal{L}(\mathbf{w}(t))] \le -\frac{\eta(1-\delta^2)}{2}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2}, \tag{17}$$

which implies

$$\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \le \frac{2\mathbb{E}[\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t+1))]}{\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2}. \tag{18}$$

By summing up for $t = 1, \dots, T$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \le \frac{2\mathbb{E}[\mathcal{L}(\mathbf{w}(1)) - \mathcal{L}(\mathbf{w}(T+1))]}{T\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2} \le \frac{2\mathcal{L}(\mathbf{w}(1))}{T\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2}. \tag{19}$$

By setting $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \le \frac{2(1+\nu)L\mathcal{L}(\mathbf{w}(1))}{\sqrt{T}(1-\delta^2)} + \frac{\sigma^2}{(1+\nu)\sqrt{T}(1-\delta^2)}. \tag{20}$$

A.2 Case II: Unbiased Stochastic Gradient

For the second case of unbiased stochastic gradient, recall that the update step of SGD is

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t) \tag{21}$$

where $\nabla\ell(\mathbf{w}(t))$ is the stochastic version of the gradient of loss function $\nabla\mathcal{L}(\mathbf{w}(t))$ at $\mathbf{w}(t)$ and $\eta > 0$ is the learning rate, and the element of $\hat{\mathbf{m}}(t)$ is given by

$$\hat{m}_j(t) = \begin{cases} \frac{1}{p_j(t)}, & \text{if } w_j(t) \in \mathcal{S}(t), \\ 0, & \text{otherwise,} \end{cases} \tag{22}$$

with $p_i(t) = \mathbb{P}(w_i(t) \in s_i(t))$.

We suppose that the stochastic gradient $\nabla\ell(\mathbf{w}(t))$ is unbiased, i.e., $\mathbb{E}[\nabla\ell(\mathbf{w}(t))] = \nabla\mathcal{L}(\mathbf{w}(t))$. Then we have $\mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)] = \mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \mathbf{p}(t)^{-1} \odot \mathbf{m}(t)] = \mathbb{E}[\mathbb{E}_{\mathbf{m}(t)}[\nabla\ell(\mathbf{w}(t)) \odot \mathbf{p}(t)^{-1} \odot \mathbf{m}(t)|\nabla\ell(\mathbf{w}(t))]] = \mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \mathbf{p}(t)^{-1} \odot \mathbb{E}_{\mathbf{m}(t)}[\mathbf{m}(t)|\nabla\ell(\mathbf{w}(t))]] = \mathbb{E}[\nabla\ell(\mathbf{w}(t))] = \nabla\mathcal{L}(\mathbf{w}(t))$, indicating that $\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)$ is also unbiased. We make the following assumption for the stochastic gradient $\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)$.

Assumption 4 (Bounded Variance). We assume that the stochastic gradient $\nabla\ell(\mathbf{w}) \odot \hat{\mathbf{m}}(t)$ is biased and its variance is bounded. That is, for any $\mathbf{w}(t)$ and $\hat{\mathbf{m}}(t)$, we have

$$\mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)] = \nabla\mathcal{L}(\mathbf{w}(t)), \tag{23}$$

and

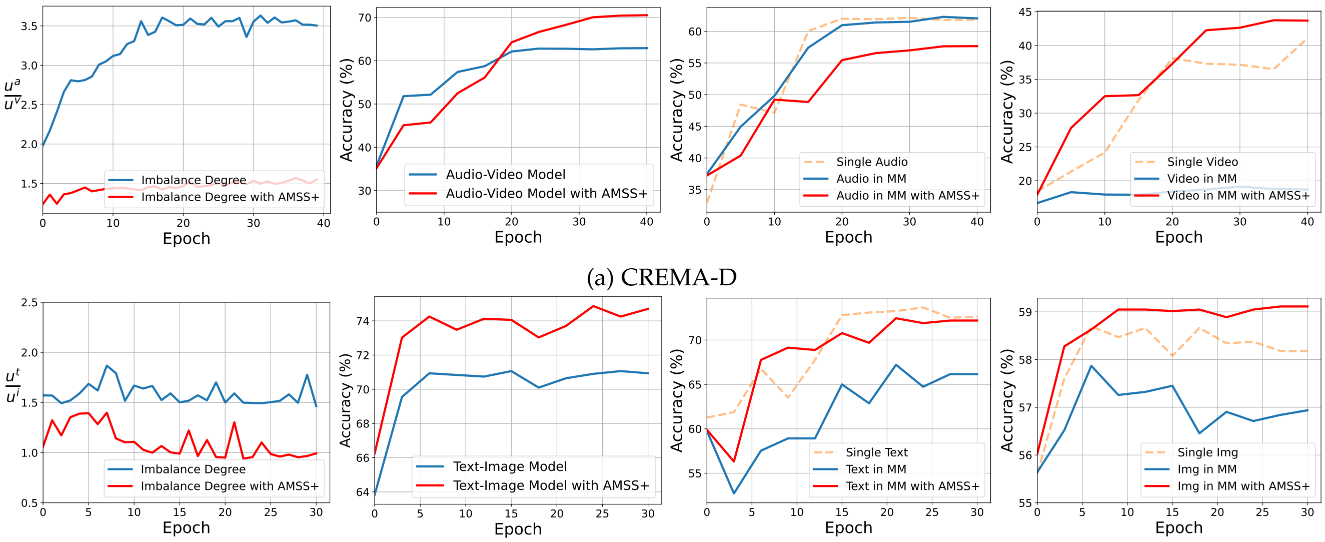
$$\mathbb{E}[\|\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t) - \mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)]\|]^2 \le \nu\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \sigma^2, \tag{24}$$

where $\sigma^2 \ge 0$ and $\nu \ge 0$ are two constants.

We present the convergence property in the following theorem.

Theorem 4 (Formal, AMSS+). Under Assumptions 1, 4, if the learning rate is set as $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$, then

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \le \frac{2(1+\nu)L\mathcal{L}(\mathbf{w}(1))}{\sqrt{T}} + \frac{\sigma^2}{(1+\nu)\sqrt{T}}. \tag{25}$$



(a) CREMA-D

(b) Twitter-15

Fig. A: The same as Figure 5, but for CREMA-D and Twitter-15 datasets.

Proof 2. By Assumption 1, we have

$$\begin{aligned}\mathcal{L}\left(\mathbf{w}(t+1)\right)-\mathcal{L}\left(\mathbf{w}(t)\right)\leq&\langle\nabla\mathcal{L}\left(\mathbf{w}(t)\right),\mathbf{w}(t+1)-\mathbf{w}(t)\rangle+\frac{L}{2}\|\mathbf{w}(t+1)-\mathbf{w}(t)\|^2\\ \overset{(21)}{=}&-\eta\langle\nabla\mathcal{L}\left(\mathbf{w}(t)\right),\hat{\mathbf{g}}(t)\rangle+\frac{\eta^2L}{2}\|\hat{\mathbf{g}}(t)\|^2\end{aligned}\tag{26}$$

where $\hat{\mathbf{g}}(t) := \nabla\ell\left(\mathbf{w}(t)\right)\odot\hat{\mathbf{m}}(t)$. Taking expectation over both sides of (26) and by using Assumption 4, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t+1)\right)-\mathcal{L}\left(\mathbf{w}(t)\right)]\leq&-\eta\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta^2L}{2}\mathbb{E}[\|\hat{\mathbf{g}}(t)\|^2]\\ =&-\eta\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta^2L}{2}(\mathbb{E}[\|\hat{\mathbf{g}}(t)-\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\|^2]+\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\|^2)\\ \leq&-\eta\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta^2L}{2}\left((1+\nu)\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2+\sigma^2\right)\\ \leq&-\eta\|\nabla\mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2+\frac{\eta^2L\sigma^2}{2},\end{aligned}\tag{27}$$

where the last inequality is due to $\eta\leq\frac{1}{(1+\nu)L}$. Therefore, we have

$$\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2\leq\frac{2\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t)\right)-\mathcal{L}\left(\mathbf{w}(t+1)\right)]}{\eta}+\eta L\sigma^2.\tag{28}$$

By summing up for $t=1,\dots,T$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2\leq\frac{2\mathbb{E}[\mathcal{L}\left(\mathbf{w}(1)\right)-\mathcal{L}\left(\mathbf{w}(T+1)\right)]}{T\eta}+\eta L\sigma^2\leq\frac{2\mathcal{L}\left(\mathbf{w}(1)\right)}{T\eta}+\eta L\sigma^2.\tag{29}$$

By setting $\eta=\frac{1}{(1+\nu)L\sqrt{T}}$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2\leq\frac{2(1+\nu)L\mathcal{L}\left(\mathbf{w}(1)\right)}{\sqrt{T}}+\frac{\sigma^2}{(1+\nu)\sqrt{T}}.\tag{30}$$

Appendix B
Additional Experiment Analysis

B.1 Training Overhead Analysis

To demonstrate the efficiency of our AMSS/AMMS+ in practice, we analyze the algorithm complexity and conduct an experiment for training overhead comparison in this section.

Specifically, we first analyze the algorithm complexity. Compared with naive MML approach, our methods introduce the additional computation of mask vector, i.e., $\mathbf{m}(t)$. The complexity for computing mask vector is $\mathcal{O}(|\mathbf{w}|)$. Then, we conduct an experiment to demonstrate the efficiency of AMSS/AMSS+. We adopt standard multimodal learning method which utilizes naive joint training strategy, and two state-of-the-art baselines PMR [18], and AGM [25] for comparison. PMR introduces additional prototype cross-entropy and prototypical entropy regularization losses on a single modality using a prototype model, while AGM requires multiple inputs to adjust gradient update weights for each modality, both of which significantly increase training overhead. Furthermore, all experiments are conducted on a workstation with an Intel(R) Xeon(R) Gold 5220R CPU and an A6000 GPU. For all methods, we use the same number of epochs and batch size on each dataset.

The accuracy and training overhead on all datasets are reported in Table I. From Table I, we can see that: 1). Compared to the Naive method, AMSS/AMSS+ can achieve significantly better accuracy with comparable training overhead. 2). Compared with PMR and AGM, AMSS/AMSS+ can achieve better performance in iterms of accuracy and training overhead.

TABLE I: Comparison of training overhead. The best and the second-best results are shown in **bold** and underline, respectively.

Metric	Dataset	Method				
		Naive	PMR	AGM	AMSS	AMSS+
Training Overhead (minutes)	Kinetics-Sound	**120.02**	349.32	153.10	123.66	123.70
	CREMA-D	**90.62**	180.66	117.68	92.45	92.46
	Sarcasm-Detection	**199.23**	433.94	244.36	210.11	210.21
	Twitter-2015	**40.21**	91.23	51.63	42.02	42.02
	NvGesture	**502.88**	-	696.46	535.47	535.58
Accuracy	Kinetics-Sound	64.55%	66.56%	66.02%	68.96%	**72.25%**
	CREMA-D	63.31%	66.59%	67.07%	67.61%	**70.30%**
	Sarcasm-Detection	82.86%	83.60%	84.02%	84.14%	**84.35%**
	Twitter-2015	70.11%	74.25%	74.83%	**75.89%**	75.12%
	NvGesture	78.63%	-	80.71%	82.57%	**84.64%**

B.2 Analysis of Modality Imbalance

Figure A illustrates the analysis of modality imbalance in two additional datasets (CREMA-D, Twitter-15). Similar to the results depicted in Figure 5, the modality imbalance degree on both datasets exhibits a reduction through the AMSS+ strategy. On the CREMA-D dataset, our method initially performs less favorably than the Baseline method but exhibits significant improvement in later stages compared to the Baseline. Analysis of the curve of imbalance degree variations indicates that early application of the AMSS+ modulation strategy mitigates the dominance of the Audio modality optimization, enabling the model to better explore information from Video modality and ultimately achieve superior performance. Furthermore, on the Twitter15 dataset, the balancing strategy implemented by AMSS+ results in improved performance across both modality branches and overall model performance. This improvement is characterized by achieving performance levels for individual modality branches in multimodal models that are comparable to those achieved through separate training of single modality models. Additional experiments once again validate the effectiveness and reliability of AMSS+.

Appendix C
Implementation Details

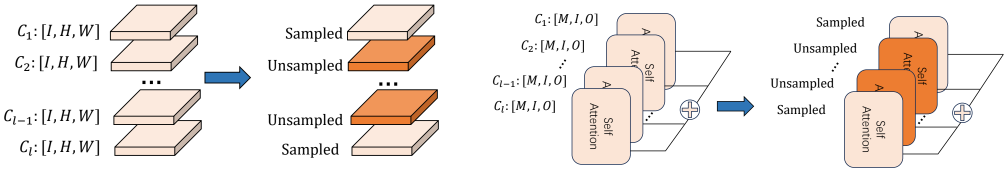
C.1 Optimization details

We use stochastic gradient descent (SGD) as the optimizer for the audio-video and NVGesture datasets, with a momentum of 0.9 and weight decay of 1e-4. The initial learning rate is set to 1e-2, and when the loss is saturated, it is multiplied by 0.1. For the text-image dataset, following [58, 59], we use Adam as the optimizer, with an initial learning rate of 1e-5. In the context of the audio-video datasets, the scaling factor τ is configured to be 0.25, while on the text-image and NVGesture datasets, the τ is set to 0.5. We train all models on a single RTX 4090 GPU.

C.2 Implementation Details for Different Model Architecture

To further improve the training efficiency for AMSS/AMMS+, we design a channel-wise mask unit for CNN and a head-wise mask unit for Transformer.

More specifically, for CNN, we sum the importance scores of the remaining dimensions along the channel dimension (*channel-wise mask unit*) to determine which channels should be masked. We suppose the dimensions of each CNN layer as $[O, I, H, W]$, where O, I, H, W denote the number of output channels, the number of input channels, height and width of the convolutional kernel, respectively. Then, we use the output channels as mask units, where $C_l = O$. As shown in Figure Ba, after calculating the importance of all parameters, we sum the importance values of all parameters



(a) The l-th convolutional layer in CNN

(b) The l-th MHA Layer in Transformer

Fig. B: The implemented details of subnetwork selection for different model architectures. C_l represents the number of mask units in the l-th layer of the network. “Sampled” indicates the mask units to be updated, while “Unsampled” refers to those that are not updated.

corresponding to each output channel across dimensions I, H, and W to assess the significance of that channel. Based on the given sampling ratio ρ, we sample across all channels. For instance, when the sampling ratio is 50%, we select $\frac{O}{2}$ channels according to the proposed sampling method and mask the remaining parameters, preventing them from participating in the update process.

For Transformer, a similar strategy is applied to the model along the head dimension, i.e., head-wise mask unit. Specifically, we consider the dimensions of multi-head attention as $[N, M, I, O]$, where N represents the number of attention heads, M denotes the number of matrices, I is the input dimension, and O is the output dimension. Then, we treat the number of attention heads as the mask units, i.e., $C_l = N$. As illustrated in Figure Bb, we sum the importance values of all parameters across dimensions M, I, and O to assess the significance of each attention head. Subsequently, based on a specified sampling ratio ρ, we conduct importance sampling across all attention heads. For instance, with a sampling ratio of 50%, we select $\frac{N}{2}$ attention heads using the proposed sampling method and mask the remaining heads.

Additionally, the scope of subnetworks $\mathcal{S}$ varies depending on the fusion methods used, as the parameters for each modality differ based on the chosen fusion strategy. Specifically, in late fusion, the classifiers are considered part of the modality parameters, so each modality subnetwork includes both its classifier and encoder. In other fusion scenarios, each modality subnetwork consists only of its encoder.