Supplementary Material for "Learning to Rebalance Multi-Modal Optimization by Adaptively Masking Subnetworks"

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

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

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

$$m_j(t) = \begin{cases} 1, & \text{if } w_j(t) \in \mathcal{S}(t), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

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

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}') \leq \langle \nabla \mathcal{L}(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{L}{2} \|\mathbf{w} - \mathbf{w}'\|^2. \tag{3}$$

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

$$\mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)] = \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)), \tag{4}$$

and

$$\mathbb{E}[\|\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t) - \mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)]\|]^2 \leq \nu \|\nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\|^2 + \sigma^2, \tag{5}$$

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

$$\|\mathbb{E}[\nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t) - \nabla \ell(\mathbf{w}(t))]\| \leq \delta \|\mathbb{E}[\nabla \ell(\mathbf{w}(t))]\|, \tag{6}$$

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

We present the convergence property in the following theorem.

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

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2(1+\nu)L\mathcal{L}(\mathbf{w}(1))}{\sqrt{T}(1-\delta^2)} + \frac{\sigma^2}{(1+\nu)\sqrt{T}(1-\delta^2)}. \tag{7}$$

Proof 1. By Assumption 1, we have

$$\begin{aligned} \mathcal{L}(\mathbf{w}(t+1)) - \mathcal{L}(\mathbf{w}(t)) \leq & \langle \nabla \mathcal{L}(\mathbf{w}(t)), \mathbf{w}(t+1) - \mathbf{w}(t) \rangle + \frac{L}{2} \|\mathbf{w}(t+1) - \mathbf{w}(t)\|^2 \\ \overset{(1)}{=} & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \mathbf{g}(t) \rangle + \frac{\eta^2 L}{2} \|\mathbf{g}(t)\|^2 \end{aligned} \tag{8}$$

where $\mathbf{g}(t) := \nabla \ell(\mathbf{w}(t)) \odot \mathbf{m}(t)$. Taking expectation over both sides of (8) and by using Assumption 2, we have

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\mathbf{w}(t+1)) - \mathcal{L}(\mathbf{w}(t))] \leq & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta^2 L}{2} \mathbb{E}[\|\mathbf{g}(t)\|^2] \\ = & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta^2 L}{2} (\mathbb{E}[\|\mathbf{g}(t) - \mathbb{E}[\mathbf{g}(t)]\|^2] + \mathbb{E}[\|\mathbb{E}[\mathbf{g}(t)]\|^2]) \\ \leq & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta^2 L}{2} \left((1+\nu) \|\nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\|^2 + \sigma^2 \right) \\ \leq & -\eta \langle \nabla \mathcal{L}(\mathbf{w}(t)), \nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t)) \rangle + \frac{\eta}{2} \|\nabla \mathcal{L}(\mathbf{w}(t)) + b(\mathbf{w}(t))\|^2 + \frac{\eta^2 L \sigma^2}{2}, \end{aligned} \tag{9}$$

where the last inequality is due to $\eta \leq \frac{1}{(1+\nu)L}$. Since $-\langle a, b\rangle + \frac{\|b\|^2}{2} = \frac{\|a-b\|^2}{2} - \frac{\|a\|^2}{2}$, then (9) implies that

$$\begin{aligned}\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t+1)\right) - \mathcal{L}\left(\mathbf{w}(t)\right)] \leq & -\eta\langle \nabla\mathcal{L}\left(\mathbf{w}(t)\right), \nabla\mathcal{L}\left(\mathbf{w}(t)\right) + b(\mathbf{w}(t))\rangle + \frac{\eta^2 L}{2}\mathbb{E}[\|\mathbf{g}(t)\|^2] \\ \leq & \frac{\eta}{2}\|b(\mathbf{w}(t))\|^2 - \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2}, \end{aligned} \tag{10}$$

Next, by (4) in Assumption 2 and (6) in Assumption 3, we know

$$\|b(\mathbf{w}(t))\| = \|\mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \mathbf{m}(t) - \nabla\ell(\mathbf{w}(t))]\| \leq \delta\|\mathbb{E}[\nabla\ell(\mathbf{w}(t))]\| = \delta\|\nabla\mathcal{L}(\mathbf{w}(t))\|. \tag{11}$$

Therefore, by (10) and (11) we have

$$\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t+1)\right) - \mathcal{L}\left(\mathbf{w}(t)\right)] \leq -\frac{\eta(1-\delta^2)}{2}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2}, \tag{12}$$

which implies

$$\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t)\right) - \mathcal{L}\left(\mathbf{w}(t+1)\right)]}{\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2}. \tag{13}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2\mathbb{E}[\mathcal{L}\left(\mathbf{w}(1)\right) - \mathcal{L}\left(\mathbf{w}(T+1)\right)]}{T\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2} \leq \frac{2\mathcal{L}\left(\mathbf{w}(1)\right)}{T\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2}. \tag{14}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2(1+\nu)L\mathcal{L}\left(\mathbf{w}(1)\right)}{\sqrt{T}(1-\delta^2)} + \frac{\sigma^2}{(1+\nu)\sqrt{T}(1-\delta^2)}. \tag{15}$$

A.2 Case II: Unbiased Stochastic Gradient

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

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t) \tag{16}$$

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

$$\hat{m}_j(t) = \begin{cases} \frac{1}{p_j(t)}, & \text{if } w_j(t) \in \mathcal{S}(t), \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

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

$$\mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)] = \nabla\mathcal{L}(\mathbf{w}(t)), \tag{18}$$

and

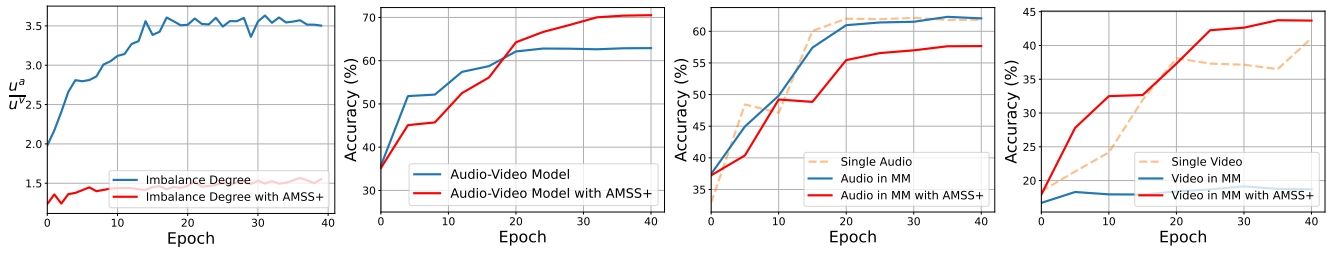
$$\mathbb{E}[\|\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t) - \mathbb{E}[\nabla\ell(\mathbf{w}(t)) \odot \hat{\mathbf{m}}(t)]\|]^2 \leq \nu\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 + \sigma^2, \tag{19}$$

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

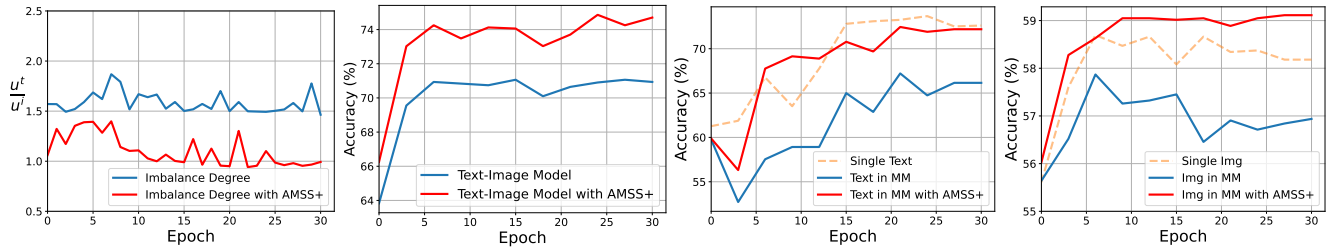
We present the convergence property in the following theorem.

Theorem 2 (Formal, AMSS+). Under Assumptions 1, 4, if the learning rate is set as $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$, then

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



(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

$$\begin{aligned}\mathcal{L}\left(\mathbf{w}(t+1)\right)-\mathcal{L}\left(\mathbf{w}(t)\right) \leq & \langle \nabla \mathcal{L}\left(\mathbf{w}(t)\right), \mathbf{w}(t+1)-\mathbf{w}(t)\rangle+\frac{L}{2}\|\mathbf{w}(t+1)-\mathbf{w}(t)\|^2 \\ \overset{(16)}{=} & -\eta\langle \nabla \mathcal{L}\left(\mathbf{w}(t)\right), \hat{\mathbf{g}}(t)\rangle+\frac{\eta^2 L}{2}\|\hat{\mathbf{g}}(t)\|^2 \end{aligned} \tag{21}$$

where $\hat{\mathbf{g}}(t) := \nabla \ell\left(\mathbf{w}(t)\right) \odot \hat{\mathbf{m}}(t)$. Taking expectation over both sides of (21) and by using Assumption 4, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t+1)\right)-\mathcal{L}\left(\mathbf{w}(t)\right)] \leq & -\eta\|\nabla \mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta^2 L}{2}\mathbb{E}[\|\hat{\mathbf{g}}(t)\|^2] \\ = & -\eta\|\nabla \mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta^2 L}{2}(\mathbb{E}[\|\hat{\mathbf{g}}(t)-\nabla \mathcal{L}\left(\mathbf{w}(t)\right)\|^2]+\|\nabla \mathcal{L}\left(\mathbf{w}(t)\right)\|^2) \\ \leq & -\eta\|\nabla \mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta^2 L}{2}\left((1+\nu)\|\nabla \mathcal{L}(\mathbf{w}(t))\|^2+\sigma^2\right) \\ \leq & -\eta\|\nabla \mathcal{L}\left(\mathbf{w}(t)\right)\|^2+\frac{\eta}{2}\|\nabla \mathcal{L}(\mathbf{w}(t))\|^2+\frac{\eta^2 L\sigma^2}{2}, \end{aligned} \tag{22}$$

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

$$\|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2\mathbb{E}[\mathcal{L}\left(\mathbf{w}(t)\right)-\mathcal{L}\left(\mathbf{w}(t+1)\right)]}{\eta}+\eta L\sigma^2. \tag{23}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla \mathcal{L}(\mathbf{w}(t))\|^2 \leq \frac{2\mathbb{E}[\mathcal{L}\left(\mathbf{w}(1)\right)-\mathcal{L}\left(\mathbf{w}(T+1)\right)]}{T\eta}+\eta L\sigma^2 \leq \frac{2\mathcal{L}\left(\mathbf{w}(1)\right)}{T\eta}+\eta L\sigma^2. \tag{24}$$

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

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

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 [1], and AGM [2] 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 A. From Table A, 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 A: 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%**

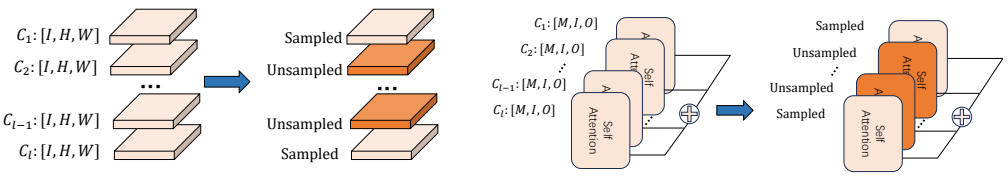
B.2 Analysis of Modality Imbalance

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+.

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 [3, 4], 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.



(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.

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 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.

REFERENCES

[1] 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.

[2] 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.

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

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