

## A Appendix

### A.1 Dataset

To evaluate the effectiveness of our approach, we conducted experiments on three widely used multimodal datasets from the domains of affective computing and multimedia. From the affective computing domain, we utilized the CREMA-D and UR-Funny datasets.

Table 3: The datasets specifications.

Field of Research	Size	Dataset	Modality	Samples	content
Affective Computing	L	UR-Funny[11]	{a, v, t}	16,514	humor
	M	CREMA-D [2]	{a, v}	7,442	emotion
Multimedia	M	AV-MNIST[32]	{a, v}	70,000	digit
	S	AVE[30]	{a, v}	4,143	event detection

The CREMA-D dataset, curated for speech emotion recognition tasks, includes six emotional labels spanning various speech recordings. The UR-Funny dataset, designed for humor detection, incorporates multimodal cues, including text, visual gestures, and acoustic-prosodic features, providing a rich benchmark for affective computing. In the multimedia domain, we employed the AV-MNIST dataset, which focuses on multimedia classification. This dataset features disturbed images paired with audio signals, offering a challenging setting for evaluating cross-modal learning.

Further details about these datasets, including modality types and task descriptions, are provided in Table. 3.

### A.2 Mono-Modal Contribution in Total Performance

To split the total loss function into the modality-specific loss function as what Eq. 4 shows, we adopted the Shapley metric proposed by [22]. Consider:

$$\Phi(x) = f^s(\phi_1, \phi_2, \dots, \phi_M; \theta^s), \quad (13)$$

where  $x = (x_1, \dots, x_M)$  represents all corresponding  $M$  modalities of the input and  $\phi_m = f_m^e(x_1^i, \dots, x_M^i; \theta_m^e, \theta_{m1}^s, \dots, \theta_{mk}^s \mid_{m \neq k}, \dots, \theta_{mM}^s)$  is defined as the feature representation produced by the  $m$ -th encoder taking into account other modalities' effect through the fusion network's parameters,  $\theta_{mj}$ . Lets  $\mathcal{M} := \{m\}_{i=1}^M$  denote the set of all modalities. Zero-padding  $0_m$  indicates the absence of modality  $m$  features. If  $S$  is a subset of  $\mathcal{M}$ , then  $\Phi(S)$  indicates that for  $m \in S$ ,  $x_m$  is replaced with  $0_m$ . The mono-modal response for modality  $m$  is then defined as:

$$\Phi_m(x) = \sum_{\substack{S \subseteq \mathcal{M} \setminus \{m\} \\ S \neq \emptyset}} \frac{|S|! (k - |S| - 1)!}{k!} V_m(S; \Phi), \quad (14)$$

where  $V_m(S; \Phi) = \Phi(S \cup \{m\}) - \Phi(S)$ . The empty subset is excluded from the summation to ensure the relation:

$$\Phi(x) = \sum_m \Phi_m(x). \quad (15)$$

For illustration, in the case of two modalities, it would be as follows;

$$\Phi_{m_1}(x) = \frac{1}{2} [\Phi(\{x_1, x_2\}) - \Phi(\{0_1, x_2\})] + \Phi(\{m_1, 0_{m_2}\}). \quad (16)$$

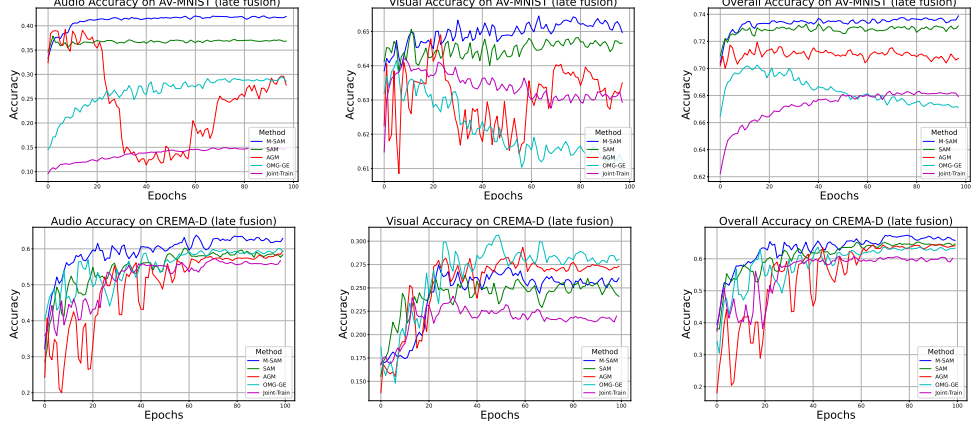


Figure 4: **Performance comparisons of late fusion settings:** Joint-Trained, AGM, OGM-GE, SAM, and our proposed Modality-Aware SAM, on the AV-MNIST and CREMA-D datasets. Viewing this in color is recommended.

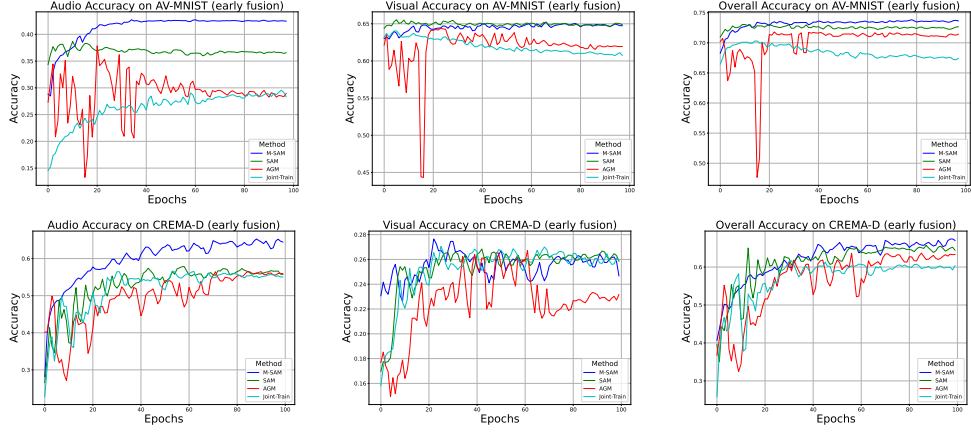


Figure 5: **Performance comparisons of early fusion settings:** Joint-Trained, AGM, OGM-GE, SAM, and our proposed Modality-Aware SAM, on the AV-MNIST and CREMA-D datasets. Viewing this in color is recommended.

## 772 A.2.1 From Mono-Modal Attribute to Loss Landscape Decomposition

773 The Shapley values in Eq. 14 serve as decomposing weight to compute mono-modal contributions  
 774 into the overall accuracy. They inherently operate in the domain of performance metrics, not loss  
 775 values. In contrast, SAM is explicitly designed to operate on the optimization landscape by directly  
 776 utilizing loss values. This fundamental difference raises a critical challenge: how can we bridge  
 777 the gap between modality-specific performance contributions and the loss landscape decomposition  
 778 during the training process?

779 To address this, we propose leveraging the Shapley value in Eq. 14. By computing Shapley values  
 780 for each modality during training, we dynamically adjust the weights associated with their losses,  
 781 thereby linking the concept of modality performance to loss decomposition. This dynamic adjustment  
 782 ensures that modalities contributing more to performance are appropriately emphasized in the loss  
 783 computation. Specifically, the weight associated with each modality’s loss is derived from its  
 784 normalized Shapley value. Mathematically, this is expressed as:

$$\nu_m = \frac{\Phi_m}{\sum_{i=1}^M \Phi_i}, \quad (17)$$

where  $M$  is the total number of modalities,  $\Phi_m$  is the Shapley value of the  $m$ -th modality, and  $\nu_m$  represents the normalized weight in Eq. 4.

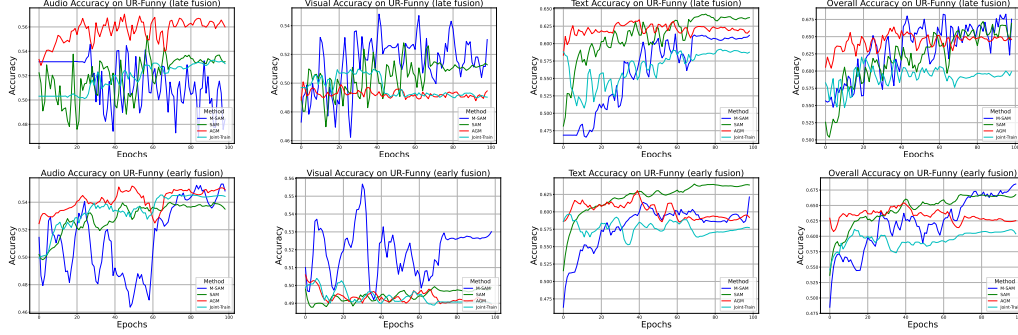


Figure 6: Performance comparisons of the Joint-Trained, AGM, SAM, and our M-SAM on UR-Funny datasets' validation set using *late fusion* (first row) and *early fusion* (second row) architecture. It is recommended to view this in color.

### A.3 Learning Curves

Fig. 4 and Fig. 6 present the performance comparisons of our proposed Modality-Aware SAM (M-SAM) with SAM and other state-of-the-art methods, including AGM, OMG-GE, and Joint-Train, under late fusion settings for the AV-MNIST and CREMA-D datasets and both late and early fusion scenarios of UR-Funny dataset, respectively. The results consistently demonstrate the superiority of M-SAM across both datasets and all metrics (audio, visual, and overall accuracy).

For the AV-MNIST dataset, M-SAM achieves the highest accuracy in all metrics. In the audio modality, M-SAM converges quickly and achieves a final accuracy of approximately 0.40, outperforming SAM (0.35), AGM, and OMG-GE, which lag significantly. In the visual modality, M-SAM demonstrates smooth and steady improvement, achieving the best accuracy of around 0.65. In contrast, SAM achieves slightly lower accuracy, and AGM and OMG-GE show notable instability. Regarding overall accuracy, M-SAM reaches approximately 0.74, clearly surpassing SAM (0.71) and significantly outperforming other methods, which fail to approach competitive levels. On the CREMA-D dataset, M-SAM maintains its superiority across all metrics. In the audio modality, M-SAM consistently achieves higher accuracy and improved stability compared to SAM, which shows oscillations during training. AGM and OMG-GE perform poorly, failing to converge effectively. In the visual modality, M-SAM achieves the highest accuracy, with SAM trailing behind and other methods struggling to maintain stability. Finally, in overall accuracy, M-SAM again emerges as the best-performing method, with the smoothest convergence and highest final accuracy.

### A.4 Margin of Superiority over Joint-Train baseline

To evaluate the effectiveness of each method, we report the normalized marginal improvement in accuracy over the Joint Training (JT) baseline. Specifically, we compute the percentage increase in overall accuracy ( $Acc_{mm}$ ) for each method relative to the JT baseline using the following formulation:

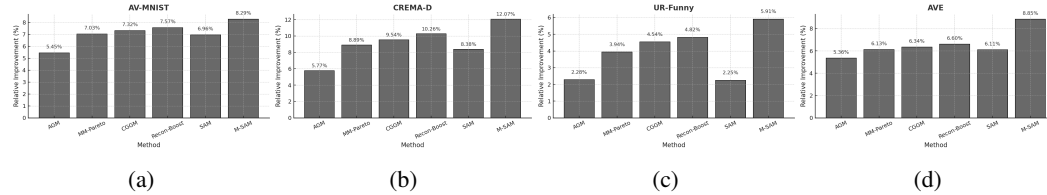


Figure 7: Relative improvement in multi-modal accuracy ( $Acc_{mm}$ ) over the Joint-Training baseline on (a) AV-MNIST, (b) CREMA-D, (c) UR-Funny, and (d) AVE datasets. Our M-SAM consistently achieves the highest normalized gain, outperforming all methods.

810 This normalized metric allows for a fair comparison across datasets of varying base difficulty and  
 811 scales, highlighting how much each method improves over standard joint optimization.

$$\Delta_{\text{rel}}(\text{Method}) = \frac{\text{Acc}_{mm}(\text{Method}) - \text{Acc}_{mm}(\text{JT})}{\text{Acc}_{mm}(\text{JT})} \times 100 \quad (18)$$

812 As shown in Figure 7, our proposed *M-SAM* consistently achieves the largest relative gain across all  
 813 datasets, outperforming strong baselines such as Recon-Boost and CGGM.