Adaptive-Margin Masking and Restoration for Balanced Multimodal Learning

Guangbin Zhang **Zheyang Luo** **Jiangming Liu***
School of Information Science and Engineering, Yunnan University, China
{zhanggb928,luozheyang}@stu.ynu.edu.cn, jiangmingliu@ynu.edu.cn

Abstract

Real-world tasks rely on information from multiple sensory sources, motivating multimodal learning as a core paradigm in modern machine learning. However, multimodal learning suffers from modality laziness, where one modality suppresses the contribution of the other modalities by dominating the training process. Recent studies of proposing various metrics to identify lazy modalities and manipulate optimizations to mitigate this imbalance, facing two main problems: 1) The identification of lazy modality is incomplete and overly sharp; 2) The dominated modality can be lagged by under-optimized lazy modalities. To address these issues, we propose **A**daptive-Margin **M**asking and **Re**storation (**AMRe**), which introduces adaptive-margin modality identification and restoration optimization to balance dominated and lazy modalities. Experimental results show that AMRe consistently outperforms competitive baselines and several state-of-the-art methods by achieving significant improvement on standard multimodal benchmarks. The codes are released in `https://github.com/LangNexus-Group/AMRe`.

1 Introduction

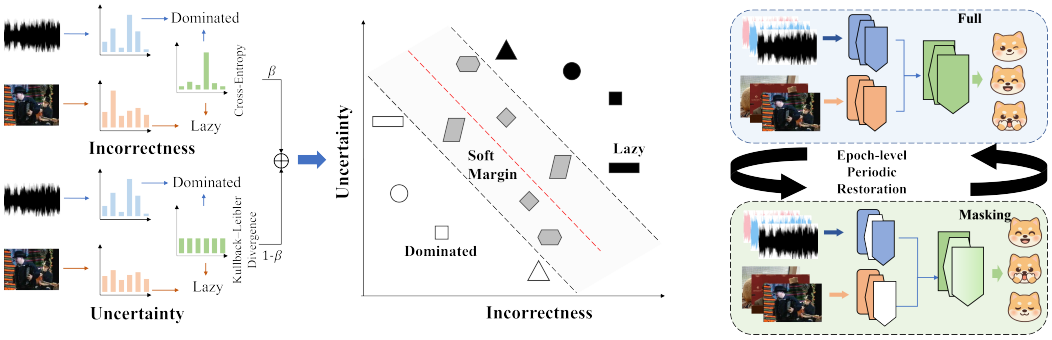
Models trained on unimodal data fail to solve the real-world tasks that require complementary and multisensory information [1, 2, 3]. Multimodal learning, which trains models on data from diverse modalities, has emerged as a central research paradigm in modern machine learning [4, 5, 6, 7]. However, recent multimodal learning methods struggle to effectively digest knowledge across modalities, because features from one modality across instances can dominate the model training. As a result, modalities fail to contribute equally, leading to the problem of modality laziness [8, 9, 10].

To address this problem, several methods have been proposed that can be broadly categorized into two directions: joint optimization and alternating optimization. Joint optimization regulates the training dynamics by tailoring learning rates [11, 12] and gradient updates [13, 14, 15, 16] based on the performance of individual modalities. Alternating optimization mitigates the suppression of lazy modalities through modality-level interventions. The effectiveness of these methods hinges on the accurate identification of the lazy modality. Representative approaches employ diverse metrics, including sample-level marginal contributions [17] and uncertainty-based measures such as probability distribution divergence [18].

However, existing lazy-modality identifications are incomplete and overly-sharp. Depending on a single metric may lead to misjudging a modality with low incorrectness but high uncertainty (or vice versa) as dominated, thereby hindering the gradient updates for the actual lazy modality. We argue that the lazy modality should be identified by *incorrectness* and *uncertainty*. As shown in the left of Figure 1 (a), incorrectness captures the modality tendency of incorrect prediction by random, while uncertainty reflects unreliability of its confidence. The other problem is overly-sharp identification.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Adaptive-Margin Modality Identification

(b) Restoration Strategy

Figure 1: (a) Adaptive-Margin Modality Identification, where left is quantified modality laziness via incorrectness and uncertainty, and right is a soft margin used to distinguish dominated (white) vs. lazy (black) modality by considering ambiguous samples (gray). (b) Restoration Strategy: The framework epoch-level periodically alternates between full optimization and adaptive masking. Masking (white) decouples the gradient flow of dominated modalities to force the exploitation of lazy ones. Shapes denote distinct instances, while colors represent different modalities.

As shown in the right of Figure 1 (a), the black shapes of high uncertainty and incorrectness are lazy-modality samples, while the white shapes are dominated-modality samples that dominate model training. The existing methods draw a hard decision boundary using various metrics to separate these samples. However, the modalities of some samples (gray shapes) near the boundary are not as lazy as the black shapes, and treating them as such introduces noise into the optimization. The ambiguity in laziness identification can mislead balanced multimodal learning.

To alleviate these problems, we propose Adaptive-Margin Masking and Restoration (AMRe) for balanced multimodal learning. We quantify the contribution of each modality by using a hybrid metric that integrates both incorrectness and uncertainty. Based on the metric, we take an *adaptive-margin modality identification* with threshold γ to softly identify dominated and lazy modalities, as shown in the right of Figure 1 (a). During training, the dominated modality is masked to promote optimization of the lazy one. However, prolonged suppression in continuous masking blocks gradient updates, leading to feature degradation. To mitigate this effect, AMRe introduces *restoration* that periodically alternates between full and masking optimization. As shown in Figure 1 (b), this periodic reactivation enables stable co-adaptation between dominated and lazy modalities.

Our contributions are summarized as follows:

- We propose an adaptive-margin modality identification, which quantifies the contribution of each modality via incorrectness and uncertainty to address modality laziness.
- We propose a unified paradigm that combines the sample-level precision of alternating optimization with the simplicity of joint training. By dynamically masking the dominated modality, AMRe decouples gradient interference without complex multi-stage pipelines.
- The experimental results show that our proposed training methods can significantly outperform baselines and existing methods by achieving the SOTA on standard benchmarks.

2 Related Work

To mitigate modality laziness caused by imbalanced learning dynamics [8, 19], existing optimization strategies can be broadly categorized into joint optimization and alternating optimization.

Joint Optimization Joint optimization strategies maintain simultaneous inputs from all modalities while regulating the optimization process through objective adjustments and speed control. One line of work focuses on recalibrating optimization objectives to combat laziness [8, 20, 21, 10]. Wang [8] adjusts the weighting of modality losses according to unimodal overfitting. The other line of work aligns optimization dynamics by manipulating gradients and learning rates [12, 13, 22, 16, 23].

OPM [23] performs on-the-fly prediction modulation by randomly dropping features of the dominated modality according to a dynamic probability. AMRe performs deterministic masking on the dominated modality while employing a soft margin to protect ambiguous samples from unnecessarily masked.

Alternating Optimization Alternating optimization seeks to prevent the dominated modality from interfering with the lazy one by altering training sequences or masking multimodal inputs, thereby isolating unimodal branches to enhance their feature representation capabilities [19, 17, 18, 24]. Wei et al. [17] introduce sample-level marginal contributions to distinguish between dominated and lazy modalities, employing resampling to boost the expressiveness of lazy modalities. Ma et al. [18] utilize KL divergence to quantify sample uncertainty, identifying modality strength to guide a data remixing strategy for targeted training of lazy samples. AMST[24] adopts a skipping strategy to reduce modality dominance, its control remains coarse-grained, relying on a fixed epoch-level freezing rather than sample-adaptive regulation to handle fine-grained instance discrepancies. AMRe eliminates gradient dominance through sample-level identification, while incorporating soft margins and restoration strategies to maintain the robustness of multimodal training.

3 Preliminary

We adopt the standard paradigm of multimodal classification and uncover the underlying mechanics that trigger modality laziness through a detailed gradient analysis.

3.1 Multimodal Learning for Classification

We take a multimodal classification task with M modalities (e.g., audio, visual, and text). Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ is the dataset consisting of N samples. For the i-th sample $x_i = \{x_i^j\}_{j=1}^{M}$, and $y_i \in \{1, \ldots, C\}$ is the ground-truth label for C categories.

For modality j, its input is processed by the corresponding encoder $z_i^j = \varphi^j(x_i^j, \theta_j) \in \mathbb{R}^{d_{\varphi^j}}$, where θ_j represents the parameters of encoder φ^j, and d_{φ^j} is output dimension. The multimodal features are concatenated to obtain a joint representation $z_i^f = [z_i^1; z_i^2; \ldots; z_i^M]$. To enable modality-wise analysis, the weight matrix of the multimodal classifier ψ_f is decomposed by modality as $W = [W^1; W^2; \ldots; W^M]$, where each $W^j \in \mathbb{R}^{C \times d_{\varphi^j}}$, so that the logit output can be written as:

$$\hat{y}_i^f = \psi_f(z_i^f) = W z_i^f + b = W^1 \cdot z_i^1 + W^2 \cdot z_i^2 + \cdots + W^M \cdot z_i^M + b. \tag{1}$$

The corresponding multimodal cross-entropy loss is $\mathcal{L}^f = \frac{1}{|B|} \sum_{i=1}^{|B|} \mathcal{L}_{\text{CE}}(\hat{y}_i^f, y_i)$, where $|B|$ denotes the batch size, and $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss. To enforce unimodal representational capability, we introduce separate classification heads ψ_j for each modality j, producing predictions as $\hat{y}_i^j = \psi_j(z_i^j)$. The unimodal loss for modality j is computed as $\mathcal{L}^j = \frac{1}{|B|} \sum_{i=1}^{|B|} \mathcal{L}_{\text{CE}}(\hat{y}_i^j, y_i)$.

The overall training objective is defined as:

$$\mathcal{L}_{\text{std}} = \mathcal{L}^f + \alpha \cdot \sum_{j=1}^{M} \mathcal{L}^j, \tag{2}$$

where α controls the contribution of the unimodal losses.

3.2 The Modality Laziness Problem

During backpropagation, for a given sample i and modality $m \in \{1, 2, \ldots, M\}$, the gradient of the multimodal loss with respect to the parameters θ_m can be decomposed as:

$$g_{\theta_m}^{\text{multi}} = \frac{\partial \mathcal{L}_i^f}{\partial \psi_f(z_i^f)} \cdot \frac{\partial \psi_f(z_i^f)}{\partial z_i^f} \cdot \frac{\partial z_i^f}{\partial z_i^m} \cdot \frac{\partial z_i^m}{\partial \theta_m}, \tag{3}$$

where z_i^f drives a single multimodal cross-entropy loss that propagates gradients to each unimodal branch via the shared fusion pathway. While this joint optimization integrates multimodal information, it also couples the gradients received by different unimodal branches.

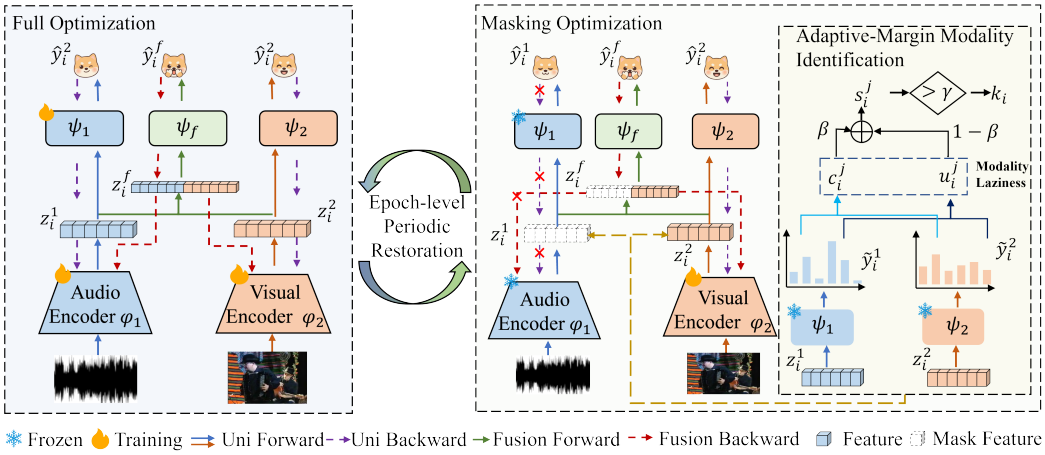


Figure 2: **Overview of AMRe. (Left)** Full Optimization, where all modalities are jointly optimized to maintain representation quality. **(Right)** Masking Optimization, in which the Adaptive-Margin Modality Identification module computes a laziness score from incorrectness and uncertainty to identify the dominated modality.

To investigate the underlying cause of this imbalance, we analyze the gradient with respect to the encoder parameters θ_m, the gradient can be formulated as

$$g_{\theta_m}^{\text{multi}} = \left[\sum_{h=1}^{C}(\text{softmax}(\hat{y}_i^f)_h - \mathbb{I}(h = y_i)) \cdot W_h^m\right] \cdot \frac{\partial z_i^m}{\partial \theta_m}, \tag{4}$$

where $\mathbb{I}(h = y_i)$ denotes the one-hot indicator of the ground-truth class. To explicitly reveal the gradient interference from the dominated modality, we expand $\text{softmax}(\hat{y}_i^f)_h$ as follows:

$$\text{softmax}(\hat{y}_i^f)_h = \frac{e^{(W_h^1 \cdot z_i^1 + \cdots + W_h^m \cdot z_i^m + \cdots + W_h^M \cdot z_i^M + b_h)}}{\sum_{r=1}^{C} e^{(W_r^1 \cdot z_i^1 + \cdots + W_r^m \cdot z_i^m + \cdots + W_r^M \cdot z_i^M + b_r)}} = \frac{e^{(W_h^m \cdot z_i^m + b_h)} P_{i,h}^m}{\sum_{r=1}^{C} (e^{(W_r^m \cdot z_i^m + b_r)} P_{i,r}^m)}, \tag{5}$$

where $P_{i,h}^m = \prod_{j \neq m}^{M} \rho_{i,h}^j$ represents the **interference term** from all other modalities, and $\rho_{i,h}^j = e^{W_h^j \cdot z_i^j}$ is defined as the **interaction factor** introduced by modality j.

When identifies a dominated modality m^* and masks its features to zero ($z_i^{m^*} = 0$), its corresponding interaction factor evaluates to $\rho_{i,h}^{m^*} = e^{W_h^{m^*} \cdot 0} = 1$. This active sample-level masking operation physically cuts off the gradient magnitude passed from the dominated modality m^* to the lazy modality m. Consequently, the updated interference term becomes $\tilde{P}_{i,h}^m = \prod_{j \notin \{m, m^*\}}^{M} \rho_{i,h}^j$, and Eq. 5 for optimizing modality m updates to:

$$\text{softmax}(\hat{y}_i^f)_h = \frac{e^{(W_h^m \cdot z_i^m + b_h)} \tilde{P}_{i,h}^m}{\sum_{r=1}^{C} e^{(W_r^m \cdot z_i^m + b_r)} \tilde{P}_{i,r}^m}. \tag{6}$$

Particularly, in a standard dual-modality scenario ($M = 2$), masking the dominated modality results in $\tilde{P}_{i,h}^m = \mathbf{1}$. Under this condition, Eq. 6 gracefully degenerates into a pure unimodal computation. This theoretical analysis directly motivates our proposed strategy: actively suppressing the dominated modality at the sample level during training, aiming to mitigate the optimization imbalance.

4 AMRe

We introduce the framework of Adaptive-Margin Masking and Restoration, which is shown in Figure 2. The core of the framework is adaptive-margin masking optimization. Together with full optimization, we introduce a periodic restoration to mitigate feature degradation.

4.1 Full Optimization

Full optimization serves as the foundational optimization strategy. It aggregates features from all modalities $\{z_i^j\}_{j=1}^M$ to construct a joint representation z_i^f for the final prediction. The objective is to minimize the task-specific loss (Eq. 2) using the complete multimodal input. This approach enables simultaneous gradient updates across all unimodal branches, fostering the learning of cross-modal interactions. We periodically reintegrate this full optimization as a restoration mechanism to recover unimodal feature expressiveness following the masking phases.

4.2 Masking Optimization

Masking optimization masks the dominated modality that is identified using adaptive-margin modality identification, while the modules corresponding to the lazy modality are trained.

Modality Laziness We propose a comprehensive metric to quantify the modality laziness by integrating sample-level *incorrectness* and *uncertainty* with a soft margin.

- **Incorrectness:** We quantify unimodal classification performance for incorrectness using the cross-entropy loss:

$$c_i^j = \mathcal{L}_{\mathrm{CE}}(\tilde{y}_i^j, y_i), \tag{7}$$

  where $\tilde{y}_i^j$ is the prediction probability obtained from the frozen classification head of modality j. A larger score c_i^j indicates that modality j provides less discriminative information for sample x_i under the current training state. This suggests that modality j may be under-optimized relative to other modalities and may behave as a lazy modality.
- **Uncertainty:** We employ the negative Kullback–Leibler (KL) divergence to measure the deviation of the modality prediction probability distribution $\tilde{y}_i^j$ from a uniform distribution U for *uncertainty*:

$$u_i^j = -D_{\mathrm{KL}}(\tilde{y}_i^j \| U), \tag{8}$$

  where u_i^j quantifies the information gain provided by a modality beyond random guessing. A larger u_i^j indicates the prediction approaches uniformity, indicating higher uncertainty and reduced reliability.

We formulate the laziness score by integrating incorrectness and uncertainty:

$$s_i^j = \beta \cdot \sigma_\tau(c_i^j) + (1-\beta) \cdot \sigma_\tau(u_i^j), \tag{9}$$

where $\sigma_\tau(\cdot)$ is the temperature-scaled softmax normalization, which projects the raw metrics into a normalized probability distribution to ensure commensurability across modalities. We set the temperature $\tau = 0.2$ to control the sharpness of the score distribution. The coefficient $\beta \in [0, 1]$ governs the trade-off between incorrectness and uncertainty.

Adaptive-Margin Modality Identification For sample x_i, $\mathcal{S}_i = \{s_i^1, \ldots, s_i^M\}$ is the set of laziness scores across all modalities. We quantify the optimization imbalance by computing the laziness scores gap Δ_i:

$$\Delta_i = \max(\mathcal{S}_i) - \min(\mathcal{S}_i). \tag{10}$$

To determine the masking target, we first identify the dominated modality index $m_i^* = \arg\min(\mathcal{S}_i)$, which exhibits the lowest laziness score. To avoid masking on ambiguous samples, where the discrepancy in laziness scores is marginal, we introduce a soft margin γ. The adaptive mask index k_i is formally defined as:

$$k_i = \begin{cases} m_i^*, & \text{if } \Delta_i \geq \gamma, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where $k_i \in \{0, 1, \ldots, M\}$ denotes the index of the modality to be masked for the i-th sample. This formulation ensures that the dominated modality i is selected to be masked ($k_i = m_i^*$) if $\Delta_i \geq \gamma$, and all modalities are kept for full optimization ($k_i = 0$) if $\Delta_i < \gamma$.

Fusion Feature with Masked Modality We apply the masking operation to the feature embeddings before fusion. If modality j is selected to be masked (i.e., $k_i = j$), its feature vector z_i^j is replaced by a zero vector $\mathbf{0}$; otherwise, the original feature is retained. The joint representation z_i^f is obtained by concatenating these mask-conditioned features. This active zero-masking cuts off the gradient flow from the dominated modality to the lazy modality encoders within the fusion loss, thereby forcing the optimization of the lazy modality.

4.3 Training Strategy

We propose a restoration that periodically reverts to full optimization, which is shown in Figure 2. Unlike continuous masking where prolonged suppression leads to optimization lag, this strategy prevents representation degradation by regularly enabling full gradient updates. After an initial warm-up phase of E_{WU} epochs, the optimization alternates between full and masking optimization based on a restoration interval E_p. This process is regulated by a binary indicator:

$$I_{\text{mask}} = \mathbb{I}(e \bmod E_p \neq 1 \wedge e > E_{WU}), \tag{12}$$

where $e \in \{1, 2, \ldots, E\}$ denotes the current epoch and E is the total number of training epochs. Specifically, $I_{\text{mask}} = 0$ triggers full joint optimization to stabilize features, while $I_{\text{mask}} = 1$ activates sample-level masking optimization to reduce gradient interference from dominated modalities.

In adaptive-margin masking, the standard unimodal loss formulation becomes inadequate, as samples with masked modalities should not propagate gradients through their corresponding unimodal branches. To address this limitation, we reformulate the unimodal loss $\mathcal{L}^j$ for modality j as:

$$\widetilde{\mathcal{L}}^j = \frac{\sum_{i=1}^{|B|} (\mathbb{I}(k_i \neq j)\, \mathcal{L}_{\text{CE}}(\hat{y}_i^j, y_i))}{\sum_{i=1}^{|B|} \mathbb{I}(k_i \neq j)}, \tag{13}$$

where $k_i \in \{0, 1, \ldots, M\}$ denotes the modality masked for sample i, and $\mathbb{I}(\cdot)$ is the indicator function. $\widetilde{\mathcal{L}}^j$ reduces to the standard unimodal loss if no modality is masked. Together with the standard fusion loss obtained, the overall training objective of AMRe is defined as:

$$\mathcal{L}_{\text{AMRe}} = \mathcal{L}^f + \alpha \cdot \sum_{j=1}^{M} \widetilde{\mathcal{L}}^j, \tag{14}$$

which jointly optimizes the multimodal fusion and the mask-conditioned unimodal objectives.

5 Experiments

Experiments are carried out on standard multimodal benchmarks, and we compare the proposed AMRe to several competitive baselines.

5.1 Experimental Setup

Datasets The benchmarks cover diverse modalities: CREMA-D [25] and AVE [26] for audio-visual tasks, MVSA-Single [27] for visual-text tasks, and IEMOCAP [28] for audio-visual-text tasks.[2]

Baselines We group existing approaches into three categories. (1) *Conventional fusion* methods include parameter-free aggregation [29], such as concatenation (Concat) and summation (Sum), as well as parameterized interaction mechanisms, including FiLM [30] and Bi-Gated [31]; (2) *Joint optimization* methods include OGM-GE [13],OPM[23], as well as LFM [3]; (3) *Alternating optimization* approaches include MLA [19], Resample [17], Remix [18], and AMST [24].[3]

Backbones ResNet-18 [32] for audio-visual, BERT [33] + MAE [34] for visual-text, and a combination of BERT (text encoder) and ResNet-18 (audio-visual encoder) for tri-modal tasks.[4]

[2]Detailed dataset statistics are provided in Appendix C.1

[3]Further details are available in Appendix C.2.

[4]Implementation details can be found in Appendix C.3.

Table 1: Results(%) across baselines with our proposed methods (AMRe). **Bold** is the best result, and underlining is the second best. * is parameterized interaction.

Method	CREMA-D	AVE	MVSA	IEMOCAP	Avg
Concat	73.52	72.89	73.18	65.38	71.24
Sum	73.66	70.40	73.39	69.38	71.71
FiLM* [30]	71.51	68.66	72.35	66.72	69.81
Bi-Gated* [31]	71.91	71.89	72.73	66.72	70.81
Concat + AMRe	**81.18**	**76.37**	**75.68**	68.93	**75.54**(+4.30)
Sum + AMRe	76.08	71.14	73.60	**71.30**	73.03(+1.32)
FiLM* + AMRe	73.39	70.40	73.80	67.31	71.23(+1.42)
Bi-Gated* + AMRe	75.13	72.39	73.60	67.90	72.25(+1.44)

Table 2: Results of accuracy (%) and training time (s) with our models and SOTA models. **Bold** is the best result, and underlining is the second best. - indicates that there are no relative modality.

Method	CREMA-D		AVE		MVSA		IEMOCAP		Avg	
	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time
Audio	63.17	18.23	64.68	13.53	-	-	48.52	26.52	-	-
Visual	68.95	18.91	37.56	9.26	60.08	18.98	28.85	26.35	48.86	18.38
Text	-	-	-	-	70.27	18.79	58.14	26.89	-	-
Concat	73.52	21.24	72.89	17.15	73.18	36.64	65.38	38.73	71.24	28.44
Random	74.33	21.73	72.14	17.56	73.18	36.94	65.53	39.02	71.30	28.81
OGM-GE [13]	73.45	22.25	72.64	17.94	74.22	37.63	65.79	40.16	71.52	29.50
LFM [3]	77.28	21.45	72.39	17.17	73.31	36.81	66.42	38.94	72.52	28.59
OPM [23]	78.76	22.58	74.38	15.45	73.39	37.68	65.83	39.83	72.99	28.89
MLA[19]	76.48	20.37	75.62	16.78	75.26	36.28	67.01	38.64	73.59	28.02
Resample[17]	75.00	35.93	71.89	22.92	74.43	56.32	66.57	66.50	71.98	45.42
Remix[18]	74.52	25.87	72.69	21.60	74.01	38.52	67.16	40.95	72.10	31.74
AMST[24]	80.50	34.66	74.38	30.93	72.60	64.65	68.40	85.41	73.97	53.91
AMRe	**81.18**	21.45	**76.37**	17.39	**75.68**	36.70	**68.93**	39.04	**75.54**	28.64

5.2 Results

AMRe with Conventional Fusion To evaluate the universality and plug-and-play nature of AMRe, we first verify its effectiveness within parameter-free aggregation paradigms. As shown in Table 1, although these methods are computationally efficient, they lack intrinsic mechanisms for balanced optimization. Both parameter-free aggregation baselines obtain significant performance gains if integrated with AMRe, achieving average improvements of 4.30% and 1.32%, respectively. Additionally, AMRe benefits parameterized interaction by achieving 1.43% on average. This suggests that AMRe effectively alleviates modality imbalance arising from simple fusion operations. We adopt Concat as the baseline fusion strategy in subsequent experiments. Parameterized fusion models underperform parameter-free aggregation in certain cases. Detailed analysis is provided in Appendix B.1.

Comparison with SOTA models As shown in Table 2, AMRe is evaluated against a broad set of state-of-the-art baselines and consistently achieves competitive performance across all benchmarks. AMRe outperform the optimization-based approaches such as LFM, highlighting the advantage of adaptive-margin masking over global gradient modulation. AMRe surpasses recent alternating strategies, including AMST and MLA, indicating that adaptive-margin modality identification combined with periodic restoration is more effective than rigid alternating schemes for mitigating modality laziness. Furthermore, compared to existing sample-level masking methods (e.g., Random and OPM) that rely on fixed or dynamic probabilitieswhich risk diminishing the gradient signals for lazy modalitiesAMRe achieves superior performance by actively cutting off the gradient interference from the dominated modality. However, prolonged suppression of dominated modality can lead to degraded feature representations over time. AMRe implements a periodic restoration strategy that alternates between adaptive masking and full optimization at the epoch level, which helps

preserve feature quality throughout training. AMRe attains the highest average accuracy of 75.54%, demonstrating robust performance in balanced multimodal learning.

Training Time In the same experimental setup on a single V100 GPU, AMRe introduces less than 1.5% additional latency compared to the Concat baseline. This efficiency is mainly due to the lightweight metric computations ($O(C)$) relative to backbone feature extraction, and the fact that masking is skipped during the restoration phases. Detailed analysis are provided in Appendix D.1.

5.3 Ablation Study

Table 3 presents a comprehensive ablation study to examine the contributions of incorrectness, uncertainty, soft margin, and restoration interval. Hyperparameter analysis is provided in Appendix D.2.

Restoration Interval Comparing the variant without restoration to the full model underscores the importance of periodic restoration. When restoration is disabled, continuous masking keeps one modality suppressed for extended periods, which can lead to degraded features. This effect is most pronounced on AVE, where performance even falls below the baseline. Reintroducing the restoration phase periodically resumes full optimization, counteracting this degradation and yielding consistent improvements. We discussed different recovery strategies in Appendix D.3.

Table 3: Ablation study on AMRe. We investigate the impact of incorrectness (c), uncertainty (u), the soft margin (γ) and the restoration interval (E_p).

c	u	γ	E_{p}	CREMA	AVE	MVSA	IEMOCAP
×	×	×	×	73.52	72.89	73.18	65.38
✓	✓	✓	×	77.55	71.80	73.39	67.90
✓	×	×	✓	75.13	75.16	73.80	67.01
×	✓	×	✓	<u>79.84</u>	72.64	73.18	<u>68.34</u>
✓	✓	×	✓	78.76	<u>75.37</u>	<u>74.43</u>	66.42
✓	×	✓	✓	75.94	75.12	74.01	65.83
×	✓	✓	✓	79.70	73.63	73.60	67.31
✓	✓	✓	✓	**81.18**	**76.37**	**75.68**	**68.93**

Incorrectness versus Uncertainty Uncertainty generally yields higher performance than incorrectness alone, suggesting that prediction confidence can provide a more informative signal than loss magnitude for identifying lazy modalities. However, neither cue is sufficient by itself. Combining uncertainty and incorrectness consistently yields more stable performance, highlighting their synergy in accurately characterizing laziness. This integration builds a holistic view of learning dynamics, ensuring robustness against data noise and calibration errors. To demonstrate that the superior performance explicitly stems from the combination of two distinct metrics, we provide a detailed discussion on various metric combinations in Appendix D.4.

Soft Margin This strategy yields consistent improvements, particularly on ambiguity-heavy datasets like IEMOCAP, where performance increases by 2.51% compared to the hard masking baseline. By adopting a soft margin to identify the dominated modality, the model can avoid rigid decision boundaries when modality contributions are comparable. This helps reduce unnecessary masking oscillations caused by noise and stabilizes the optimization process.

5.4 Cross-method controlled ablation

We conduct cross-method controlled experiments by modifying one component at a time while keeping the remaining method-specific settings unchanged.

Masking Policy To isolate the effect of the masking policy, we replace only the masking operator while keeping the remaining components of each method unchanged. For OPM, we retain its original batch-level discriminative diagnosis and replace Bernoulli feature dropping with deterministic masking. For AMRe, we retain its sample-level

Table 4: Accuracy (%) under different masking policies.

Method	CREMA-D	AVE
OPM w/ probabilistic dropping	78.76	74.38
OPM w/ deterministic masking	80.11	75.37
AMRe w/ probabilistic dropping	77.28	73.88
AMRe w/ deterministic masking	**81.18**	**76.37**

diagnosis, soft margin, and restoration schedule, and replace deterministic masking with probabilistic dropping. As shown in Table 4, deterministic masking consistently improves both OPM and AMRe on CREMA-D and AVE. This result indicates that deterministic masking provides a more targeted intervention than probabilistic dropping. Notably, AMRe still outperforms OPM when both adopt deterministic masking, suggesting that its performance gain is not solely attributable to the masking operator.

Soft Margin To isolate the effect of the soft margin, we introduce the same no-intervention rule into Remix, OPM, and AMRe while retaining each method's original modality-discrepancy criterion and balancing operation. Specifically, when the native modality discrepancy is below the margin γ, the original balancing intervention is skipped. As shown in Table 5, introducing the soft margin improves all three methods on both CREMA-D and IEMOCAP. The consistent improvements across different balancing approaches indicate that the benefit of the soft margin is not specific to AMRe. By introducing a no-intervention region for samples with similar modality contributions, the soft margin reduces unnecessary balancing operations caused by ambiguous modality relationships.

Table 5: Accuracy (%) of different methods with and without the soft margin γ.

Method	CREMA-D	IEMOCAP
Concat	73.52	65.38
Remix	74.52	67.16
Remix w/ γ	77.55	67.46
OPM	78.76	65.83
OPM w/ γ	79.43	66.51
AMRe w/o γ	78.76	66.42
AMRe	**81.18**	**68.93**

Restoration To isolate periodic restoration, we apply the same restoration strategy across methods while retaining their original balancing mechanisms, periodically suspending the balancing operation for full multimodal optimization. As shown in Table 6, periodic restoration improves all evaluated methods on both CREMA-D and IEMOCAP. These consistent gains suggest that periodically returning to full-modality optimization helps mitigate prolonged suppression of the dominated modality and allows all modality branches to realign with the joint multimodal objective. The effectiveness of restoration across different balancing methods further indicates that its benefit is not specific to AMRe.

Table 6: Accuracy (%) of different methods with and without periodic restoration (rest.).

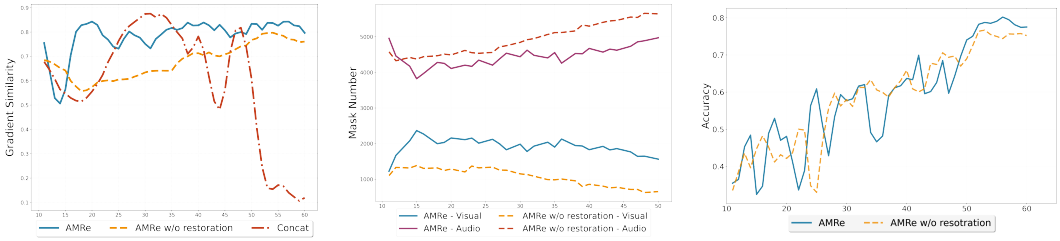
Method	CREMA-D	IEMOCAP
Concat	73.52	65.38
Remix	74.52	67.16
Remix w/ rest.	76.61	**70.71**
Resample	75.00	66.57
Resample w/ rest.	76.08	67.46
OPM	78.76	65.83
OPM w/ rest.	79.81	67.57
AMRe w/o rest.	77.55	67.90
AMRe	**81.18**	68.93

5.5 Restoration

We conduct a systematic analysis of the restoration, as shown in Figure 3, focusing on representation stability, masking frequency, and capability upper bound.

Stability As shown in Figure 3(a), unimodal gradient similarity over time highlights the issue of feature degradation under different optimization strategies. Concat shows a clear decline in training stages, indicating an increasing optimization imbalance as modality learning diverges from the joint objective. AMRe w/o restoration mitigates this sharp decline but converges to a relatively low gradient similarity, suggesting that prolonged suppression by masking dominated modality limits the alignment between modalities. AMRe maintains consistently high gradient similarity throughout training. By periodically reintroducing full optimization, it stabilizes gradient interactions and alignments.

Masking Frequency Figure 3(b) shows the frequency of masking dominated modality across different strategies. Under AMRe w/o restoration, the Audio modality is relatively stronger and is therefore masked much more frequently than the Visual modality. This frequent masking reduces



(a) Modality Gradient Similarity (b) Mask Frequency (c) Fusion Modality Accuracy

Figure 3: Visualization of dynamics optimization across different strategies, where the horizontal axis represents training epochs.

the sample coverage of effective Audio updates rather than indicating that Audio becomes weaker. Periodic restoration reactivates the masked modality under full multimodal optimization, helping recover its alignment with the joint objective and yielding a more balanced masking distribution.

Capability Upper Bound As shown in Figure 3(c), AMRe exhibits smoother convergence and higher final accuracy than the version without restoration. Although masking optimization can address the laziness of the modality, it weakens the dominated modality. Restoration recovers the contribution of the dominated modality by alternating training between full and masking optimizations. The periodic fluctuations arise from this alternation: masking redirects optimization toward the lazy modality, while restoration reactivates all branches and realigns the fusion model, ultimately leading to better final performance.

5.6 Applicability to MLLMs

As shown in Table 7, integrating the AMRe mechanism into the Ola-7B framework consistently improves overall multimodal performance (Mul.) on both the AVE and MVSA datasets, regardless of the fusion strategy. Examining the unimodal results reveals that these gains are primarily driven by the activation of lazy modalities. In the AVE dataset, for instance, the visual modality is notably weaker than the audio modality. With AMRe, the visual accuracy under the Concat strategy increases from 64.68% to 70.56%. A similar trend is observed in the MVSA dataset: the relatively weak textual modality improves from 66.11% to 70.06% under the Concat strategy. Detailed experimental setup is provided in the Appendix D.7.

Table 7: Results(%) of different methods on Ola-7B.

Method	**AVE**			**MVSA**		
	Aud.	Vis.	Mul.	Tex.	Img.	Mul.
Concat	84.33	64.68	90.55	66.11	**68.40**	74.43
Concat + AMRe	86.57	**70.56**	**91.79**	**70.06**	68.19	**75.68**
Sum	86.82	56.21	89.30	66.94	65.28	73.39
Sum + AMRe	**87.31**	60.70	91.04	67.15	66.11	73.80

6 Conclusion

We present Adaptive-Margin Masking and Restoration, a novel approach to address the challenge of modality laziness for balanced multimodal learning. AMRe proposes to identify laziness of modality with a soft margin, considering incorrectness and uncertainty of modality. Based on identification, we effectively adopt the feature of lazy modalities by masking the dominated modality to propose masking optimization. Additionally, together with restoration, AMRe balances the multimodal training avoiding the masked dominated modality to be lagged by the over-training of lazy modality. The experimental results with insight analysis show that the proposed model outperforms the state-of-the-art models on different benchmarks.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (grant No. 62566064) and in part by Yunnan Fundamental Research Projects (grant No. 202401CF070189).

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[2] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+ x: A survey on multimodal learning in the light of data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9102–9122, 2024.

[3] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems*, 37:62108–62122, 2024.

[4] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1407–1417, 2021.

[5] Shicai Wei, Yang Luo, Xiaoguang Ma, Peng Ren, and Chunbo Luo. MSH-Net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[6] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36: 76749–76771, 2023.

[7] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[8] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

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

[10] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pages 8632–8656. PMLR, 2023.

[11] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[12] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021.

[13] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[14] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023.

[15] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86. Springer, 2024.

[16] Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22879–22888, 2025.

[17] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27338–27347, 2024.

[18] Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. *arXiv preprint arXiv:2506.11550*, 2025.

[19] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27456–27466, 2024.

[20] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023.

[21] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[22] Konstantinos Kontras, Christos Chatzichristos, Matthew Blaschko, and Maarten De Vos. Improving multimodal learning with multi-loss gradient modulation. *arXiv preprint arXiv:2405.07930*, 2024.

[23] Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2024.

[24] Hugo Manuel Alves Henriques e Silva, Hongguang Chen, and Selpi. AMST: Alternating Multimodal Skip Training. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 526–541. Springer, 2025.

[25] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[26] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018.

[27] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, pages 15–27. Springer, 2016.

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

[29] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018.

[30] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[31] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[34] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[35] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020.

[36] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[37] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

A The Algorithm of AMRe

Algorithm 1 AMRe Training Strategy

Inputs: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, encoders φ_j, fusion head ψ_f, unimodal head ψ_j for j-th modality, restoration interval E_p, training epoch E, warm-up epoch E_{WU}.
for $e = 1$ **to** E **do**
  // Condition
  $I_{\text{mask}} \leftarrow (e \mod E_p \neq 1 \wedge e > E_{WU})$
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    Extract features z^j for all $j \in \{1..M\}$.
    if I_{mask} **then**
      // Adaptive-Margin Masking
      Obtain temporary predictions $\tilde{y}^j$.
      Compute Incorrectness c_i^j (Eq. 7)
      Compute Uncertainty u_i^j (Eq. 8).
      Compute Laziness Score s_i^j (Eq. 9).
      Obtain adaptive mask index k_i (Eq. 11).
    else
      // Restoration or Warm-Up
      Set $k_i \leftarrow 0$ for all samples.
    end if
    Obtain the joint representation z^f according to k.
    Obtain $\hat{y}^f \leftarrow \psi_f(z^f), \quad \hat{y}^j \leftarrow \psi_j(z^j)$.
    Update the model according to the Loss (Eq. 14).
  end for
end for

B Analysis of Parameterized Interaction

B.1 Theoretical Derivation

As shown in Figure 4, the multimodal model adopts a dual-branch architecture comprising two unimodal encoders, φ_1 and φ_2, employed to extract feature representations from their corresponding modality inputs. The outputs are $z_i^1 = \varphi_1(x_i^1; \theta_1) \in \mathbb{R}^{d_{\varphi_1}}$ and $z_i^2 = \varphi_2(x_i^2; \theta_2) \in \mathbb{R}^{d_{\varphi_2}}$, where θ_1 and θ_2 are the parameters of the unimodal encoders. We introduce specific unimodal refinement modules $\psi_{\tau_1}, \psi_{\tau_2}$ and a shared fusion module ψ_{τ_m}. The corresponding outputs are formulated as:

$$z_i^{r_1} = \psi_{\tau_1}(z_i^1; \theta_{\tau_1}), \quad z_i^{r_2} = \psi_{\tau_2}(z_i^2; \theta_{\tau_2}); \tag{15}$$

$$z_i^{\tau_1}, z_i^{\tau_2} = \psi_{\tau_m}(z_i^1, z_i^2; \theta_{\tau_m}) \tag{16}$$

These feature components are concatenated to form the joint representation $z_i^f = [z_i^{r_1}; z_i^{\tau_1}; z_i^{\tau_2}; z_i^{r_2}]$, where $[;]$ is the concatenation operation. Based on z_i^f, the prediction function is defined as:

$$\psi_f(z_i^f) = W z_i^f + b \tag{17}$$

The corresponding multimodal cross-entropy loss is defined as:

$$\mathcal{L}^f = \mathcal{L}_{CE}(\psi_f(z_i^f), y_i) \tag{18}$$

During backpropagation, the gradient of the loss with respect to the encoder parameters of the second modality, θ_2, can be decomposed using the chain rule:

$$g_{\theta_2}^{multi} = \frac{\partial \mathcal{L}^f}{\partial \psi_f(z_i^f)} \cdot \frac{\partial \psi_f(z_i^f)}{\partial z_i^f} \cdot \left(\underbrace{\frac{\partial z_i^f}{\partial z_i^{r_2}} \cdot \frac{\partial z_i^{r_2}}{\partial z_i^2}}_{\text{Direct Path}} + \underbrace{\frac{\partial z_i^f}{\partial z_i^{\tau_1}} \cdot \frac{\partial z_i^{\tau_1}}{\partial z_i^2} + \frac{\partial z_i^f}{\partial z_i^{\tau_2}} \cdot \frac{\partial z_i^{\tau_2}}{\partial z_i^2}}_{\text{Coupled Fusion Path}} \right) \cdot \frac{\partial z_i^2}{\partial \theta_2} \tag{19}$$

$$= \left[\sum_{h=1}^{C} (\text{softmax}(\psi_f(z_i^f))_h - \mathbf{I}(h = y_i)) \right] \cdot W^2 \cdot \left(\psi'_{\tau_2}(z_i^{r_2}) + \psi'^1_{\tau_m}(z_i^1, z_i^2) + \psi'^2_{\tau_m}(z_i^1, z_i^2) \right) \cdot \frac{\partial z_i^2}{\partial \theta_2}. \tag{20}$$

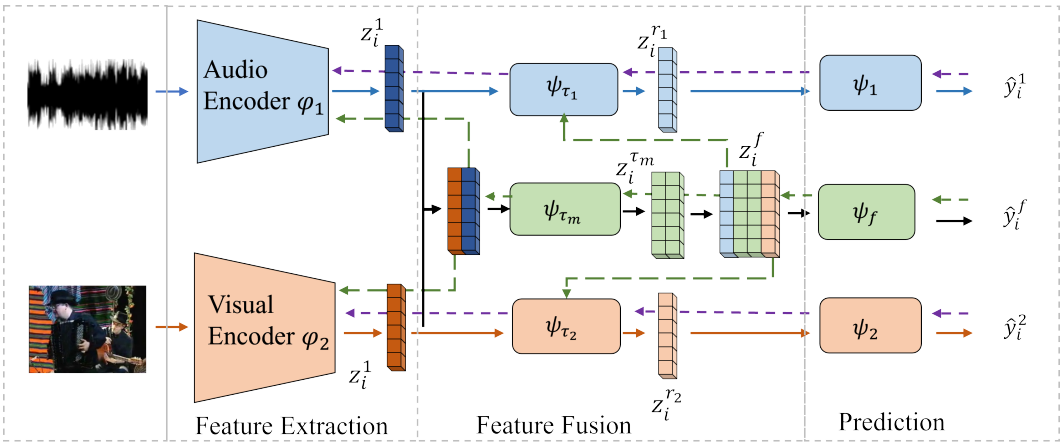


Figure 4: **Parameterized fusion optimization.** Unlike parameter-free aggregation, learnable fusion modules (ψ_{τ_m}) interact with unimodal features before prediction, introducing additional gradient coupling between modalities. Dashed lines denote gradient propagation paths.

Table 8: Results(%) of performance of AMRe combined with different fusion mechanisms. **Bold** indicates the best results.

Dataset	Concat		MMTM		CentralNet		CAEF	
	Base	+ AMRe	Base	+ AMRe	Base	+ AMRe	Base	+ AMRe
CREMA-D	73.52	**81.18**	66.40	**68.68**	66.67	**70.97**	68.95	**76.48**
AVE	72.89	**76.37**	65.67	**68.41**	69.40	**73.13**	69.90	**73.13**

This derivation reveals that the gradient update for modality 2 is influenced by modality 1 through two distinct pathways: the joint softmax calculation and the parameterized interaction module. Consequently, parameterized interaction inherently exhibits stronger gradient coupling than non-parameterized approaches, which explains its generally inferior performance when optimization imbalances occurs.

We analyze why AMRe yields smaller performance gains on parameterized interaction compared to parameter-free aggregation. Assuming modality 1 is identified as the dominated modality, we apply masking such that $z_i^1 = \mathbf{0}$. Eq. 16 then becomes:

$$z_i^{\tau_1'}, z_i^{\tau_2'} = \psi_{\tau_m}(\mathbf{0}, z_i^2; \theta_{\tau_m}) \tag{21}$$

The joint feature is updated to $z_i^f = [\mathbf{0}; z_i^{\tau_1'}; z_i^{\tau_2'}; z_i^{r_2}]$. Accordingly, the gradient in Eq. 20 is updated as:

$$g_{\theta_2}^{multi} = \left[\sum_{h=1}^{C}(\text{softmax}(\psi_f(z_i^f))_h - \mathbf{I}(h = y_i))\right] \cdot W^2 \cdot \left(\psi'_{\tau_2}(z_i^{r_2}) + \psi'^1_{\tau_m}(\mathbf{0}, z_i^2) + \psi'^2_{\tau_m}(\mathbf{0}, z_i^2)\right) \cdot \frac{\partial z_i^2}{\partial \theta_2}. \tag{22}$$

Although masking z_i^1 eliminates its direct contribution to the joint representation, the gradient flow for m_2 remains entangled with the fusion module ψ_{τ_m}. Unlike parameter-free aggregation, where masking completely decouples the modalities, parameterized interaction retains implicit coupling through the learnable parameters of ψ_{τ_m}. Even with $z_i^{m_1} = 0$, the fusion module's gradients ($\psi'^1_{\tau_m}$ and $\psi'^2_{\tau_m}$) still participate in backpropagation, exerting an implicit influence on modality 2. This residual coupling dampens the effectiveness of the explicit masking strategy, explaining why the performance boost of AMRe on parameterized interaction is observable but less pronounced than on parameter-free architectures.

B.2 Empirical Verification

To empirically verify these theoretical insights, we evaluate AMRe on three representative parameterized interaction models: MMTM [35], CentralNet [36], and CAEF (C.2). The results in Table 8 demonstrate that AMRe consistently enhances performance across diverse fusion paradigms, confirming its robustness. However, the observation that parameterized methods lag behind simple concatenation aligns with our derivation, suggesting that their inherent gradient entanglement intensifies modality imbalance and hinders the optimization of unimodal features.

Table 9: Dataset statistics for the multimodal benchmarks used in our experiments. Each dataset is split into training, validation, and test sets with an approximate ratio of 8:1:1. We report the exact number of samples in each split.

Split	CREMA-D	AVE	MVSA	IEMOCAP
Train	6,698	3,311	3,552	5,426
Val	0	402	477	683
Test	744	402	481	676
Total	**7,442**	**4,115**	**4,510**	**6,785**

C Experiment Description

C.1 Dataset

To ensure a fair comparison and align with established protocols in prior studies [19, 24], we evaluate the effectiveness of our proposed AMRe framework on four representative multimodal datasets. Most datasets are partitioned into training, validation, and testing sets with an approximate ratio of 8:1:1, except for CREMA-D, which follows a 9:1 training-testing split without a validation set, consistent with prior works. The specific statistics and partition details are summarized in Table 9. Detailed descriptions of each dataset are provided below:

CREMA-D [25] Designed for multimodal emotion recognition, CREMA-D consists of facial and vocal emotional expressions. It contains 7,442 original clips featuring actors from diverse backgrounds enacting various emotional states. The dataset covers six universal emotion categories: *Anger, Disgust, Fear, Happiness, Neutrality,* and *Sadness*.

AVE (Audio-Visual Event) [26] The AVE dataset is curated for audio-visual event localization and classification. It comprises 4,143 video clips, each lasting 10 seconds, which capture a wide range of common actions and events in daily life. The dataset spans 28 distinct event categories, requiring the model to effectively align audio and visual cues.

MVSA (Multi-View Sentiment Analysis) [27] Targeting sentiment analysis in social media content, MVSA combines textual and visual modalities to capture complex emotional expressions. The dataset consists of image-text pairs collected from Twitter. Each pair is annotated with one of three sentiment labels: *Positive, Neutral,* or *Negative*, making it a standard benchmark for image-text sentiment classification.

IEMOCAP [28] IEMOCAP is a rich multimodal resource created for emotion recognition and speech processing research. It contains approximately 12 hours of audiovisual data recorded during natural, dyadic interactions (both improvised and scripted) between actors. The dataset includes synchronized video, audio, facial motion capture, and text transcriptions. Following common evaluation protocols, we focus on five primary emotions: *Anger, Excitement, Frustration, Neutrality,* and *Sadness*.

C.2 Baselines

Concatenation Concatenation fusion integrates multimodal representations by concatenating modality-specific feature vectors along a designated dimension. A single classifier is applied to the fused representation, whose input dimension equals the sum of all encoder output dimensions, and whose output dimension corresponds to the number of target classes.

Summation Summation fusion aggregates multimodal information by independently classifying each modality and summing their unimodal predictions. Each modality-specific encoder is paired with a classifier whose input and output dimensions correspond to the encoder output and class cardinality, respectively. The summed prediction is used to compute the training loss and update all modality-specific components.

Random The random masking strategy is built upon the concatenation fusion operation. It drops modalities for each sample based on a predefined dropout probability, while strictly ensuring that at least one modality is retained per sample.

FiLM FiLM [30] Modulation introduces conditional feature modulation for multimodal fusion. One modality (e.g., audio or text) serves as a conditioning signal to modulate the representations of other modalities.

The conditioning network outputs affine parameters γ and β, which are applied to transform modality features before classification.

Bi-Gated Bilinear Gated [31] fusion models cross-modal interactions using bilinear pooling combined with a gating mechanism, explicitly capturing multiplicative correlations between modalities.

OGM OGM [13] is an early representative method for addressing modality dominance at the optimization level. It monitors the generalization discrepancy between modalities during training and dynamically downscales the gradients of the dominated modality to encourage learning from weaker modalities.

OPM OPM [23] addresses the imbalanced multimodal learning problem by targeting the feed-forward stage of optimization. It monitors the discriminative discrepancy between modalities during training and adaptively drops the feature representations of the dominated modality with a dynamic probability. By explicitly weakening the dominated modality's influence on the final prediction, OPM forces the network to rely more heavily on the suppressed, weaker modalities, thereby accelerating their optimization. Although OPM also emphasizes dominated modality masking, because it drops dominated modality features randomly with a dynamic probability, there is a possibility of dropping all modalities simultaneously. This random dropping can easily reduce the update signals for the lazy modality, thereby harming the already lazy modality.

LFM LFM [3] introduces a learnable modality gap to explicitly model optimization discrepancies between modalities. By dynamically estimating this gap, LFM adjusts the learning process to prevent any single modality from overwhelming the optimization. However, the modality gap is estimated at a global level and does not explicitly account for per-sample fusion dynamics, limiting its ability to handle localized or instance-specific modality laziness.

MLA MLA [19] mitigates modality imbalance by alternately optimizing unimodal and multimodal objectives, preventing any single modality from dominating the learning process.

Resample Resample [17] alleviates lazy modality behavior by dynamically reweighting and resampling training instances based on modality contribution, emphasizing samples where certain modalities are underutilized.

Remix Remix[18] addresses modality laziness by selectively sampling modality-specific features at the instance level and re-mixing training data to rebalance multimodal learning.

AMST AMST [24] addresses multimodal imbalance through a skipping strategy guided by modality-wise convergence rates. Modalities that converge faster are temporarily excluded from training, allowing slower modalities to catch up. Although effective in alleviating optimization imbalance, AMST makes hard modality-level decisions and lacks adaptive, sample-level control, which may lead to suboptimal handling of samples with comparable modality contributions.

MMTM MMTM[35] employs a multimodal squeeze-and-excitation mechanism to adaptively recalibrate channel-wise features across modalities.

CentralNet CentralNet[36] maintains a continuous shared central representation that interacts with unimodal features at each layer via learnable weighted aggregation.

CAEF CAEF, proposed in this work, first refines unimodal representations using self-attention and subsequently performs cross-modal fusion through a co-attention mechanism, aiming to better align modality interactions with downstream task requirements.

C.3 Experimental setup

We summarize the hyperparameter configurations for all datasets in Table 10. For a fair comparison, all methods share the same backbone architectures, data splits, optimization algorithms, and training schedules.The unified feature dimension is set to 512 or 768 depending on the backbone architecture to ensure consistent multimodal fusion.Unless otherwise specified, each experiment is conducted with the same random seed to minimize variance caused by random initialization. All experiments were implemented on a single NVIDIA V100 GPUs.

All models are optimized using AdamW, with dataset-specific learning rates and training schedules. A step-based learning rate decay is applied consistently across all benchmarks, with more aggressive decay strategies adopted for MVSA and IEMOCAP to enhance training stability. The hyperparameters γ, β, and α respectively

Table 10: Hyperparameter configurations across different datasets. BS: batch size; E: number of training epochs; LR: learning rate; WU: warm-up epochs; Step/Ratio: step-based learning rate decay schedule; WD: weight decay; Dim: unified feature dimension.

Dataset	**BS**	E	**LR**	**WU**	**Step**	**Ratio**	**WD**	γ	β	α	E_p	**Dim**
CREMA-D	64	100	1.0×10^{-3}	5	50	0.1	1.0×10^{-4}	0.15	0.1	2	4	512
AVE	64	100	1.0×10^{-3}	5	30	0.5	5.0×10^{-4}	0.05	0.7	2	2	512
MVSA	64	60	1.0×10^{-5}	0	30	0.5	5.0×10^{-4}	0.10	0.9	2	2	768
IEMOCAP	64	60	1.0×10^{-4}	0	20	0.1	1.0×10^{-4}	0.15	0.7	1	3	768

Table 11: Comparison of training time per epoch (s) across different datasets.

Dataset	Concat	OGM-GE	LFM	MLA	Resample	Remix	AMST_FULL	AMRe
CREMA-D	21.24	22.25	21.45	20.37	35.93	25.87	34.66	21.45
AVE	17.15	17.94	17.17	16.78	22.92	21.60	30.93	17.39
MVSA	36.64	37.63	36.81	36.28	56.32	38.52	64.65	36.70
IEMOCAP	38.73	40.16	38.94	38.64	66.50	40.95	85.41	39.04

control the strength of adaptive-margin masking, laziness balancing weight, and the weighting of unimodal loss terms. All three are tuned on the validation set for each dataset to achieve balanced modality contributions.

D Analysis

D.1 Computational Overhead Analysis

We analyze the computational overhead introduced by Adaptive-Margin Masking and Restoration from both empirical and theoretical perspectives, including training latency and floating-point operations (FLOPs).

Empirical Training Latency We measure the average training time per epoch (s) on a single NVIDIA V100 GPU. As shown in Table 11, AMRe incurs only a minor overhead compared to Concat:

- **Minor Overhead:** On CREMA-D and MVSA, the additional training time is 0.21s and 0.06s per epoch.
- **Largest Dataset:** On IEMOCAP, which has the highest computational demand, the overhead remains below 0.36s per epoch.

These results suggest that the incorrectness and uncertainty estimation, together with mask generation, introduce negligible latency and do not form a computational bottleneck during training.

Theoretical FLOPs Analysis The efficiency of AMRe primarily stems from operating on compact feature representations rather than raw high-dimensional inputs. The computational complexity can be decomposed as follows:

- **Backbone Cost:** The overall FLOPs are dominated by the backbone encoder (e.g., ResNet-18 or BERT), with complexity on the order of $O(L \cdot N^2 \cdot D)$.
- **AMRe Overhead:** The computation of incorrectness and uncertainty involves simple element-wise operations with complexity $O(C)$, where C denotes the number of classes. The masking operation itself is applied over feature dimensions and incurs $O(D)$ complexity.

Since $O(C) + O(D)$ is orders of magnitude smaller than the backbone cost, the additional FLOPs introduced by AMRe are negligible. Moreover, during the **restoration phase**, which is activated every E_p epochs, the masking module is bypassed entirely, resulting in zero additional FLOPs for those epochs.

Comparison with Optimization-based Methods We further compare AMRe with representative optimization-based approaches, including OGM-GE [13] and LFM [3]. As shown in Table 2, two observations support the efficiency of AMRe:

- **Advantage over Data-level Methods:** Optimization-based approaches consistently exhibit substantially lower latency than data-level balancing methods (e.g., Resample, Remix), as the latter incur significant I/O and preprocessing overhead.

Table 12: Sensitivity analysis of β and γ across four datasets. **Bold** indicates the best results.

Dataset	β Value							γ Value					
	0	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**	**0**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
CREMA-D	79.70	**81.18**	**81.18**	78.09	77.55	77.69	75.94	79.84	79.84	79.97	**81.18**	77.55	78.94
AVE	73.63	72.89	71.89	75.62	**76.37**	75.37	75.12	75.37	**76.37**	75.87	75.87	75.83	75.37
MVSA	73.60	73.18	73.18	74.01	75.47	**75.68**	74.01	74.43	74.22	**75.68**	73.60	74.64	74.84
IEMOCAP	67.31	66.12	67.46	66.42	**68.93**	65.53	65.83	66.42	67.46	65.38	**68.93**	65.68	65.38

Table 13: Impact of the restoration interval E_p (%). **Bold** denotes the best results.

E_p	**CREMA-D**	**AVE**	**MVSA**	**IEMOCAP**
0	73.52	72.89	73.18	65.38
1	77.55	71.80	73.39	67.90
2	79.84	**76.37**	**75.68**	65.24
3	79.44	73.13	73.60	**68.93**
4	**81.18**	75.62	74.22	65.58
5	79.30	73.88	73.80	67.16

- **Efficiency among Optimization Methods:** AMRe achieves comparable or better efficiency relative to OGM and LFM.
  - OGM-GE requires computing gradient norms during backpropagation, introducing additional matrix operations.
  - LFM relies on estimating learnable modality gaps or maintaining global statistics.
  - AMRe uses lightweight metric computation ($O(C)$) and simple feature masking ($O(D)$). The periodic restoration mechanism further reduces overhead by disabling masking for a substantial portion of training.

D.2 Hyperparameters

To assess the robustness of AMRe and the sensitivity of its design choices, we analyze three key hyperparameters: the balancing weight β, the soft margin threshold γ, and the restoration interval E_p.

Laziness Balancing Weight Table 12 shows the trade-off between incorrectness ($\beta \to 1$) and uncertainty ($\beta \to 0$) in identifying lazy modalities. The results reveal a distinct divergence rooted in task characteristics:

- **Uncertainty-Dominant Task (CREMA-D)** This audio-visual emotion dataset lacks explicit semantic cues and benefits most from $\beta = 0.1$, where uncertainty serves as a more reliable signal under noisy predictions.
- **Incorrectness-Dominant Tasks (IEMOCAP, AVE, MVSA)** Tasks with fast-learning or strongly informative modalities (e.g., text or salient visual objects) prefer larger β values $(0.7 - 0.9)$. In these settings, loss discrepancy offers a clearer indication of imbalance.

The choice of β is dataset-dependent and plays a critical role in identifying and mitigating modality laziness.

Soft Margin Table 12 shows that introducing a non-zero soft margin generally improves performance over hard masking ($\gamma = 0$). The preferred margin varies across datasets and appears related to data ambiguity:

- **Higher Ambiguity:** Datasets with subjective or noisy labels (e.g., CREMA-D and IEMOCAP) favor larger margins (around $\gamma = 0.15$), which help reduce unstable masking decisions near the boundary.
- **Lower Ambiguity:** Datasets with clearer semantic cues (e.g., AVE and MVSA) perform best with smaller margins ($\gamma \in [0.05, 0.10]$), where masking remains more selective.

Restoration Table 13 examines the effect of the restoration interval by comparing full optimization ($E_p = 0$), AMRe w/o restoration ($E_p = 1$), and periodic restoration ($E_p \geq 2$).

- **Inverted U-Shape Trend.** Performance follows a clear inverted U-shaped pattern as E_p increases. AMRe w/o restoration ($E_p = 1$) already improves upon the baseline by alleviating modality dominance, but extended freezing of the dominated modality eventually leads to representational degradation. Periodic restoration mitigates this effect by reintroducing full joint optimization at regular intervals, with the strongest results consistently observed for moderate intervals ($E_p \in [2, 4]$). When the interval becomes too large, performance degrades again, supporting the view that overly prolonged suppression disrupts feature balance and harms multimodal fusion.

Table 14: Results(%) of different restoration strategies.

Method	CREMA-D	AVE	MVSA	IEMOCAP
Concat	73.52	72.89	73.18	65.38
w/o Restoration	77.55	71.80	73.39	67.90
Batch-level	73.66	73.88	74.22	65.83
Epoch-level	**81.18**	**76.37**	**75.68**	**68.93**

Table 15: Result(%) of different laziness evaluation metrics. Specifically, CE Loss and Brier Score are utilized to measure incorrectness, whereas KL divergence and Entropy are employed to evaluate uncertainty. **Bold** is the best result, and underlining is the second best.

Method	CREMA-D	AVE	MVSA	IEMOCAP	AVG
Concat	73.52	72.89	73.18	65.38	71.24
Brier	76.88	73.63	74.84	66.42	72.94
CE_Loss	75.94	75.12	74.01	65.83	72.73
Entropy	80.65	74.88	71.93	66.12	73.39
KL	79.70	73.63	73.60	67.31	73.56
Brier+Entropy	77.15	75.37	74.84	65.38	73.19
Brier+KL	79.44	74.13	73.18	66.27	73.25
CE_loss+Entropy	**81.99**	76.12	74.22	67.75	75.02
CE_Loss+KL	81.18	**76.37**	**75.68**	**68.93**	**75.54**

- **Task Dependence.** The optimal interval depends on task characteristics. Ambiguity dominated datasets such as CREMA-D benefit from longer masking phases ($E_p = 4$), which encourage the model to exploit more subtle cues. In contrast, datasets with clearer semantic structure, such as AVE, require more frequent restoration ($E_p = 2$) to maintain stable and informative representations.

D.3 Restoration Strategies

To evaluate the impact of different restoration frequencies, we compare AMRe under various restoration strategies, as shown in Table 14. Specifically, "w/o Restoration" refers to continuous masking without any reversion, while "Batch-level" signifies temporarily disabling the masking operation every 10 batches to perform full optimization. "Epoch-level" represents our proposed periodic restoration at the epoch scale.

The results demonstrate that while continuous masking (w/o Restoration) significantly improves upon the Concat baseline, our Epoch-level strategy further unlocks the model's potential, achieving the highest performance across all four datasets. Notably, the Batch-level strategy performs poorly, yielding results heavily inferior to the "w/o Restoration" setting. This severe performance drop indicates that excessively frequent mode-switching within an epoch disrupts the stability of gradient updates. It prevents the network from effectively accumulating the decoupling benefits of the masking strategy and fails to stabilize the restored joint representations.

D.4 Metrics

Cross-Entropy (CE) and KL divergence are classic metrics. However, AMRe's innovation lies in conceptualizing modality laziness from two distinct, complementary dimensions: Incorrectness (result-oriented) and Uncertainty (confidence-oriented). CE and KL are merely instantiations of this concept. To prove that our gains stem from this dual-dimension philosophy rather than the specific metrics, we have supplemented experiments using alternative metrics(e.g., replacing CE with Brier Score), as shown in Table 15. The consistently superior results confirm that it is the dual-perspective framework itself that provides a more robust measurement space, rather than a coincidental combination of CE and KL.

D.5 Case Studies

As shown in Table 16, we analyze three representative samples from the AVE training phase to reveal the incomplete and overly sharp limitations of current single-metric approaches for lazy modality detection.

Table 16: Analysis of masking decisions on the AVE dataset across different sample types. "Inc" and "Unc" denote traditional single-metric masking based on incorrectness and uncertainty, respectively. m_j indicates masking modality j, while $None$ signifies no masking operation. 'm_j **Prob.**' denotes the prediction probability of modality j, and s_i^j represents the lazy score computed by Eq. 9.

Sample Type	m_1 **Prob.**	m_2 **Prob.**	s_i^1	s_i^2	**AMRe**	**Inc**	**Unc**
Ambiguous	0.8816	0.8677	0.54495	0.45505	$None$	m_2	m_2
Noise 1	0.2152	0.0355	0.70353	0.30034	m_2	m_1	m_2
Noise 2	0.0938	0.6330	0.20263	0.79737	m_1	m_2	m_2

Table 17: Results (%) of different missing modality rates on the IEMOCAP dataset. **Bold** indicates the best results.

Method	0	10%	20%	30%	40%	50%
Concat	65.38	61.98	57.99	53.70	51.04	45.86
OGM	65.98	62.28	58.43	55.18	52.12	46.60
MLA	67.01	62.42	58.73	54.69	52.22	46.01
Resample	66.57	62.87	58.73	55.33	52.51	45.97
Remix	67.16	63.46	59.76	56.66	53.99	47.78
AMST	68.40	63.43	60.73	56.89	54.12	47.62
AMRe	**68.93**	**63.90**	**61.83**	**57.36**	**54.51**	**48.04**

Ambiguous Samples In the "Ambiguous" case, both modalities exhibit high and similar prediction probabilities. While traditional methods rigidly mask m_2, AMRe's soft margin (γ) successfully detects this negligible gap and **applies no mask** ($None$), thereby preventing unnecessary intervention and allowing both healthy modalities to optimize jointly.

Noisy Samples For the first noisy sample (**Noise 1**), incorrectness-based masking misidentifies m_1 as the dominated modality, whereas uncertainty-based masking targets m_2. In contrast, AMRe's 2D metric accurately identifies m_2 as a noisy modality (characterized by high error and low uncertainty) and selectively masks it to sever its harmful gradient updates.

Conversely, for the second noisy sample (**Noise 2**), traditional methods blindly mask the dominated modality m_2 in an attempt to "force" m_1 to learn. Recognizing m_1's severe lack of informative features, AMRe avoids this pitfall; it instead masks m_1 to isolate the noise, allowing the healthy, informative modality (m_2) to correctly guide the learning process.

D.6 Robustness Analysis

To evaluate robustness across different methods, we evaluate all methods on the IEMOCAP dataset under various missing-modality rates [19]. As shown in Table 17, AMRe achieves the best performance across missing rates from 10% to 50%. In particular, compared with AMST, AMRe consistently yields higher accuracy, with a notable improvement of 1.1% at the 20% missing rate. Moreover, the performance degradation of AMRe remains relatively mild as the missing rate increases, indicating improved stability and robustness under severe modality absence.

D.7 Applicability to MLLMs

For completeness, we have presented the experimental results and discussion settings in the appendix. To verify the applicability of our method to Multimodal Large Language Models (MLLMs), we conduct evaluations on Ola-7B[37], an MLLM capable of understanding audio, visual, and textual inputs. Specifically, we freeze all backbone parameters of Ola-7B and utilize it to replace the original feature encoders in our framework, while keeping the subsequent classification head unchanged.

As shown in Table 18, integrating the AMRe mechanism into the Ola-7B framework consistently improves overall multimodal performance (Mul.) on both the AVE and MVSA datasets, regardless of the fusion strategy (Concat or Sum).Examining the unimodal results reveals that these gains are primarily driven by the activation of lazy modalities. In the AVE dataset, for instance, the visual modality is notably weaker than the audio modality. With AMRe, the visual accuracy under the Concat strategy increases from 64.68% to 70.56%. A similar trend is observed in the MVSA dataset: the relatively weak textual modality improves from 66.11%

Table 18: Results(%) of different methods on Ola-7B.**Bold** is the best result.

Method	AVE			MVSA		
	Aud.	Vis.	Mul.	Tex.	Img.	Mul.
Concat	84.33	64.68	90.55	66.11	**68.40**	74.43
Sum	86.82	56.21	89.30	66.94	65.28	73.39
Concat+AMRe	86.57	**70.56**	**91.79**	**70.06**	68.19	**75.68**
Sum+AMRe	**87.31**	60.70	91.04	67.15	66.11	73.80

to 70.06% under the Concat strategy. These results demonstrate that AMRe generalizes well to MLLMs and effectively mitigates modality dominance.

E Limitation

Although the proposed AMRe framework effectively overcomes the incomplete and overly-sharp of existing laziness identification methods, and circumvents the optimization dominance of dominated modalities via sample-level active masking, there remains room for future exploration. First, the current joint evaluation metric employs a static fusion paradigm, because our proposed AMRe framework relies on a discrete identification process that blocks all gradient flows between the modality evaluation and the fusion module. A promising future direction is to introduce a dynamic, data-driven weighting mechanism to adaptively fuse the two evaluation dimensions during training. Second, while our epoch-level periodic restoration strategy has proven empirically effective, its granularity can be further refined. Future work could investigate gradient-aware or other fine-grained adaptive mechanisms to maximize the efficacy of the restoration phase.

F Impact Statement

This work studies optimization strategies for balanced multimodal learning. While our contribution is primarily methodological, experiments are conducted on multimodal benchmarks involving affective signals (e.g., CREMA-D and IEMOCAP). Improving modality balance may enhance the robustness of multimodal models in settings where certain modalities are noisy or unreliable, potentially benefiting accessibility-related applications. As with other work involving affective data, there is a risk of misuse in surveillance or manipulative systems; however, this study focuses on the optimization of general dynamic training frameworks rather than application-specific deployments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We are confident that we have explained our contribution and scope in the abstract and introduction sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed our limitations in the section E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have provided a detailed proof of our motivation for actively masking in section 3

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup, including dataset, baseline and implementation details, was provided in section C.1, and log files are provided in the provided code repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an code repository. All datasets used in this work are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental detail was provided in section C.1. For data splits, hyperparameter selection, and the type of optimizer used, we follow the same settings as the baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report results using a fixed random seed and identical experimental settings across all methods to ensure fair comparison. Although error bars are not included, the performance gains of our method are consistent across different datasets, fusion strategies, and evaluation metrics. This consistency reduces the likelihood that the improvements are due to randomness. We will include multi-seed evaluations with standard deviation in future work.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discussed the running time in section D.1, and explained the GPU used for the experiment in section C.3.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed our impact in Sec F

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not generate any data or models potentially having a high risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite original papers, datasets and models accurately in Sec 5

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provided new assets(code) through anonymized URL.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.