CLASSIFIER-CONSTRAINED ALTERNATING TRAINING: MITIGATING MODALITY IMBALANCE IN MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modality imbalance, driven by divergent convergence dynamics across modalities, critically limits multimodal model performance. Although alternating training methods mitigate encoder-level interference, they fail to prevent dominance of classifiers by faster-converging modalities, suppressing contributions from weaker ones. To address this core limitation, we propose Classifier-Constrained Alternating Training (CCAT). Our framework first pre-trains an unbiased cross-modal classifier using bidirectional cross-attention and a regularization term that constrains modality contribution differences. This classifier is then frozen as a stable decision anchor during subsequent training, preventing bias toward any modality. To preserve modality-specific features while leveraging this anchor, we integrate modality-specific Low-Rank Adaptation (LoRA) modules into the classifier. During alternating training, CCAT updates only the encoder of the active modality and its corresponding LoRA parameters. Furthermore, a sample-level imbalance detection mechanism quantifies contribution disparities, enabling targeted optimization of severely imbalanced samples to bolster weaker modalities. Extensive experiments across multiple benchmarks demonstrate CCAT's consistent superiority: it achieves accuracy gains of +1.35% on CREMA-D, +6.76% on Kinetic-Sound and +1.92% on MVSA over state-of-the-art methods, validating the framework's efficacy in learning balanced, robust multimodal representations.

1 INTRODUCTION

Multimodal learning integrates diverse information across modalities Baltrušaitis et al. (2019); Liang et al. (2021), proving effective for numerous tasks with substantial recent progress Yang et al. (2022a); Xu et al. (2023); Yuan et al. (2025). However, such models often face persistent performance bottlenecks Wu et al. (2022) and occasionally underperform unimodal counterparts Gat et al. (2021); Yang et al. (2025). Modality imbalance Wang et al. (2020); Su et al. (2023) constitutes the root cause. Inherent inter-modal disparities in information quality and optimization induce gradient conflicts and divergent convergence speeds during training Du et al. (2023). Consequently, dominant modalities steer optimization while weaker ones are suppressed, compromising generalization capability Sun et al. (2021); Yang et al. (2024); Zhou et al. (2025b).

Existing solutions primarily co-optimize encoders and classifiers to balance modalities Xu et al. (2025), yet struggle to resolve gradient conflicts Wei & Hu (2024). To address this issue, the alternating optimization strategy has been proposed Zhang et al. (2024); Hua et al. (2024), which effectively reduces interference at the encoder level by providing each modality with independent optimization opportunities. However, they frequently overlook emergent classifier bias. Specifically, dominant modalities converge faster, steering classifier parameters toward their feature space early in training. As training progresses, even if the underperforming modalities continue to learn actively and produce substantial gradients, the classifier has already developed a structural preference for the dominant modalities. This entrenched bias hinders the effective integration of representations from weaker modalities, thus leading to the persistence of modal imbalance problems.

We empirically track modalities' average contribution Zhou et al. (2025b) throughout training as shown in Figure 1. Alternating optimization, such as Multimodal Learning with Alternating Uni-

modal Adaptation (MLA) Zhang et al. (2024), reduces initial contribution disparity (1.00 → 0.92), yet a persistent imbalance indicates entrenched classifier bias, even as encoders decouple. This confirms that encoder-level interventions alone are insufficient to resolve structural preference in classifiers. This phenomenon shares a fundamental similarity with the class imbalance problem Thrampoulidis et al. (2022), as both suffer from early-dominance-triggered bias. Majority classes skew decision boundaries initially due to numerical advantage, suppressing minority classes later. Analogously in multimodal training, dominant modalities rapidly bias the classifier through faster convergence, creating entrenched preference that persistently suppresses weaker modalities.

Inspired by class imbalance remedies that stabilize decision boundaries via fixed classifier Yang et al. (2022b), we propose Classifier-Constrained Alternating Training (CCAT). Overall, we have constructed a two-stage training framework to systematically address the issue of uneven utilization at the dataset and sample levels. First, we pretrain a shared classifier using bidirectional cross-attention attention with regularization that penalizes large discrepancies in modality contributions, yielding a relatively unbiased initial classifier. Second, within the alternating optimization framework, the shared pretrained classifier remains frozen to ensure stable optimization targets across modalities. To preserve modality-specific representational adaptation, we equip each modality with a dedicated lightweight LoRA module integrated solely on this shared classifier. Additionally, we perform secondary updates on underperforming encoders for samples with extreme modality imbalance.

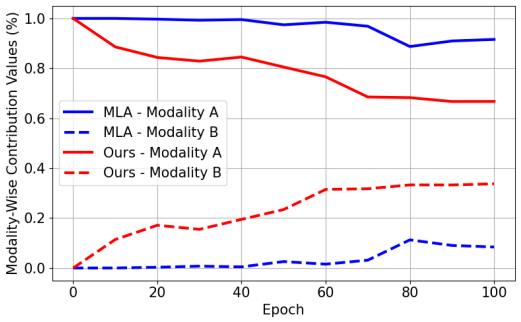


Figure 1: Evolution of modality-wise contribution values. Persistent imbalance suggests entrenched classifier preference toward dominant modalities.

Our contributions are outlined as follows: (i) Bridging class and modality imbalance through optimization dynamics, providing a new theoretical framework for understanding multimodal imbalance. (ii) Proposing CCAT with a two-stage framework that systematically addresses dataset and sample-level imbalance. (iii) Consistent SOTA improvements across three benchmarks, including over 30,000 samples. faithfully.

2 Related Work

In response to the modality imbalance problem, numerous representative methods have emerged in recent years Perez et al. (2018); Su et al. (2023); Wei et al. (2024b); Xu et al. (2025). Approaches such as On-the-fly Gradient Modulation (OGM) Peng et al. (2022), Adaptive Gradient Modulation (AGM) Li et al. (2023), and Prototypical Modality Rebalance (PMR) Fan et al. (2023) dynamically modulate modality gradients via importance measures, regulating learning rates of dominant and non-dominant modalities to balance inter-modal learning. Other approaches, including Uni-Modal Teacher (UMT) Du et al. (2023), Gradient Blending (GBlending) Wang et al. (2020), and Multi-Modal Pareto (MMPareto) Wei & Hu (2024), address objective mismatches in unimodal and multimodal learning by incorporating unimodal supervision terms into losses, enhancing weak modality representations.

However, most of the aforementioned methods primarily focus on optimizing parameter updates while overlooking intrinsic inter-modal representational disparities. To resolve this matter, several works such as Calibrating Multimodal Learning (CML) Ma et al. (2023), Multimodal BERT with Self-Distillation (MBSD) Liu et al. (2023), and LFM Yang et al. (2024), have incorporated regularization schemes, including the Kullback–Leibler (KL) divergence between unimodal predictions, modality confidence estimation, and cross-modal contrastive learning. These techniques aim to impose constraints on the modality-specific representations, thereby facilitating dynamic balancing and collaborative adaptation during training.

Further, approaches like Sample-level Modality Valuation (SMSL) Zhou et al. (2025b) and Wei et al. Wei et al. (2024a) address intrinsic data imbalance via per-sample modality contribution quantification, enhancing generalization and discriminability under imbalanced modalities.

While these methods improve cross-modal interaction efficiency, synchronous gradient updates inevitably cause conflicts between modalities Huang et al. (2022), impairing convergence stability and modality coherence. To overcome this limitation, Multimodal Learning with Alternating Unimodal Adaptation (MLA) Zhang et al. (2024) and Reconboost Hua et al. (2024) proposed a modality alternating training mechanism. This method reduces encoder level gradient interference through alternate modal encoder updates, enhancing weak modalities. However, a fundamental limitation remains. Alternating training fails to resolve disparities in modality-specific learning rates. This inherent imbalance allows dominant modalities to exert disproportionate cumulative influence on the shared classifier during early training stages, leading to structural bias, thereby perpetuating modality imbalance.

3 METHOD

3.1 SIMILARITY BETWEEN CLASS IMBALANCE AND MODALITY IMBALANCE

While modality imbalance in multimodal learning appears distinct from class imbalance in traditional machine learning, a deeper mathematical analysis reveals a fundamental connection in their gradient optimization dynamics. This section establishes a unified theoretical framework and provides a proof of their underlying similar.

For input feature $\boldsymbol{f}$ and classifier weights $\boldsymbol{W}$, the gradient of cross-entropy loss w.r.t. weight $\boldsymbol{w}_j$ is:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_j} = (\hat{y}_j - \mathbf{1}_{[j=y]})\boldsymbol{f} \tag{1}$$

where $j \in \{1, \ldots, C\}$ denotes the class index.This shows weight updates depend jointly on prediction error and feature representation.

Gradient Dynamics under Class Imbalance. In this case, the extremely low frequency of minority class samples leads the model to assign predicted probabilities $\hat{y}_j$ close to zero for them. According to the gradient formula, for minority class samples, the gradient is:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_j} \approx -\boldsymbol{f} \tag{2}$$

This approximation uncovers the core challenge of class imbalance. Although gradient directions remain theoretically valid, parameter updates become dominated by feature norm $\boldsymbol{f}$. Crucially, majority-class-dominated optimization suppresses feature magnitude in minority samples, creating a vicious cycle of feature degradation and gradient attenuation. Concurrently, inherent class imbalance diminishes the model's minority-class discriminative capacity, further impairing gradient updates.

Gradient Dynamics under Modality Imbalance. Under the multimodal learning framework, the fused feature is denoted as $\boldsymbol{f} = \gamma_1 \boldsymbol{f}^{(1)} + \gamma_2 \boldsymbol{f}^{(2)}$, where γ_1, γ_2 are not predefined fusion hyperparameters, but implicitly learned modality utilization coefficients formed during optimization. Their values reflect the classifier's degree of reliance on the feature of each modality. When modality imbalance occurs, one modality such as $\boldsymbol{f}^{(1)}$ overwhelmingly dominates due to stronger signal quality or higher data availability, resulting in $\gamma_1 \gg \gamma_2$. Then it can be approximated as:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_j} \approx (\hat{\boldsymbol{y}}_j - \mathbf{1}_{[j=y]})\gamma_1 \boldsymbol{f}^{(1)} \tag{3}$$

This gradient approximation elucidates the underlying dynamics of modality imbalance. During backpropagation, the weak-modality gradient term $(\hat{\boldsymbol{y}}_j - \mathbf{1}_{[j=y]})\gamma_2 \boldsymbol{f}^{(2)}$ is systematically suppressed in magnitude. This attenuation induces insufficient gradient signals for the weak-modality encoder, hindering effective parameter updates and causing progressive deterioration of its feature representation. Consequently, the model's estimated reliance on the weak modality diminishes, driving down the fusion weight γ_2. The reduced γ_2 further amplifies gradient suppression in subsequent updates, thereby forming a mutually inhibitory cycle of gradient attenuation and feature degradation.

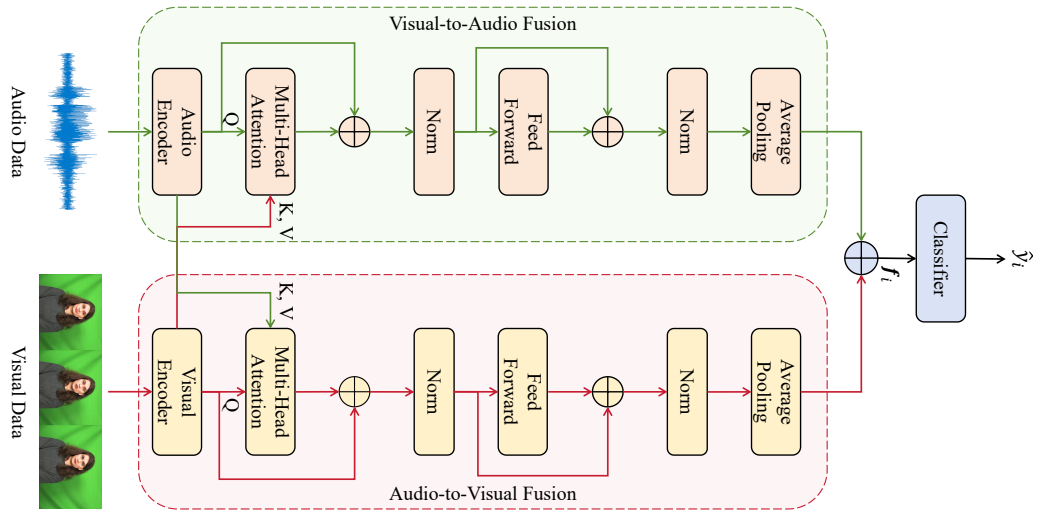


Figure 2: The fusion module using in shared classifier pretraining stage, taking the audio and video modalities as an example.

The above analysis reveals a profound theoretical isomorphism between class imbalance and modality imbalance at the gradient optimization level. Both exhibit a recursive cycle driven by early dominance bias, wherein the target component undergoes gradient suppression, representation degradation, and preference entrenchment. This bias dynamic demonstrates strong path dependence once established. Building on this insight, the implementation details of applying classifier-constraining strategies to modality imbalance will be presented in the next section.

3.2 Shared Classifier Pretraining.

Fig. 2 illustrates the overall framework of the fusion module used during the pretraining stage, taking the audio and video modalities as an example. The specific encoders employed for different modalities are described in the experimental details section. Given the complexity of multimodal tasks, we adopt a data-driven strategy to pretrain the classifier rather than relying on predefined geometric parameters. This strategy leverages a bidirectional cross-attention mechanism to dynamically fuse multimodal features (see Appendix A.1), while introducing a regularization term to constrain modality contribution disparity. This encourages the classifier to maintain an unbiased decision boundary and retain rich cross-modal interactions.

Consider a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1,2,\ldots,N}$, where each sample $\boldsymbol{x}_i = [\boldsymbol{x}_i^1, \boldsymbol{x}_i^2]$ contains two modalities $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_i^2$, accompanied by its ground-truth label y_i. Modality-specific encoders $(\mathrm{Enc}_1, \mathrm{Enc}_2)$ are employed to extract features from each modality independently:

$$\boldsymbol{z}_i^1 = \mathrm{Enc}_1(\boldsymbol{x}_i^1), \quad \boldsymbol{z}_i^2 = \mathrm{Enc}_2(\boldsymbol{x}_i^2) \tag{4}$$

The features of each modality $\boldsymbol{z}_i^1, \boldsymbol{z}_i^2$ are used as the input of the bidirectional cross-attention module $\mathrm{BiCross}(\cdot)$ for the fused features $\boldsymbol{f}_i = \mathrm{BiCross}(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2)$. This unified representation is then passed through a shared classifier $\mathrm{Cls}(\cdot)$ to generate the final prediction $\hat{y}_i$.

In order to mitigate the model's bias toward specific modalities, a modality contribution–oriented regularization mechanism is introduced. This mechanism quantifies the relative contribution of each modality to the fused representation based on the estimated mutual information (MI), which assumes statistical dependence between features of each modality $\boldsymbol{z}_i^m$ and the fused features $\boldsymbol{f}_i$. Its calculation process is shown in Figure 3 (b), and the corresponding formula is defined as follows Zhou et al. (2025b):

$$\mathrm{MI}(\boldsymbol{z}_i^m, \boldsymbol{f}_i) = \log(N) + \mathbb{E}_{\mathcal{D}}\left[\log \frac{\exp\langle \overline{\boldsymbol{f}}_i, \overline{\boldsymbol{z}}_i^m \rangle}{\sum_i \exp\langle \overline{\boldsymbol{f}}_i, \overline{\boldsymbol{z}}_i^m \rangle}\right] \tag{5}$$

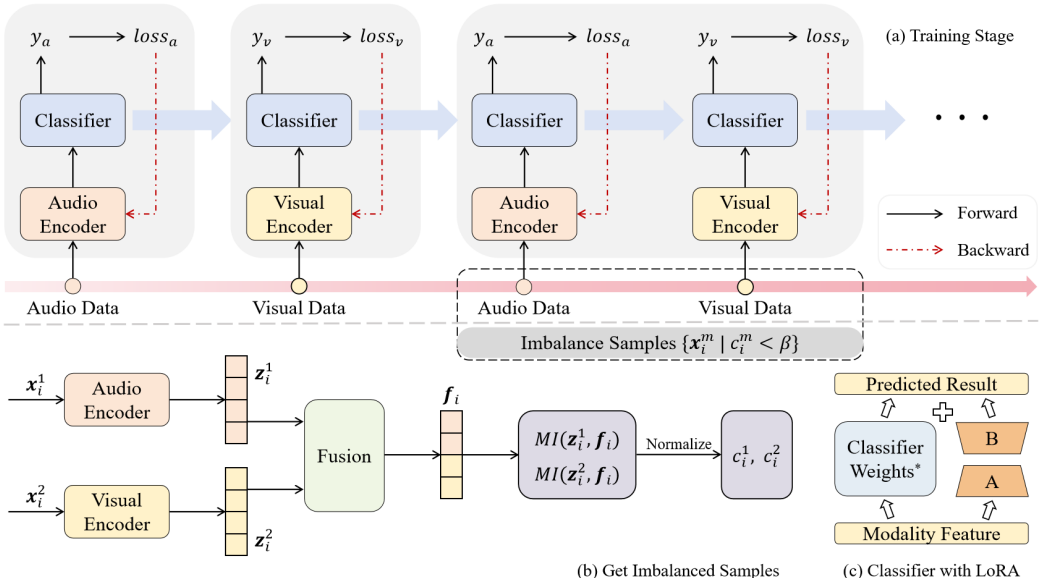


Figure 3: Overall framework employing modality-alternating training with frozen shared classifier. Per iteration, only one modality's encoder and LoRA module update using full batches. Sample-level contribution scores identify severely imbalanced cases for targeted secondary encoder and LoRA module updates. Classifier-LoRA structure detailed in (c).

where $\mathbb{E}_{\mathcal{D}}$ denotes the expectation over the dataset, and N is the total number of samples. A larger mutual information value indicates stronger influence of modality m on the fused representation.

Based on mutual information, the modality contribution vector for sample i is defined as

$$\boldsymbol{C}_i = \left[c_i^1, c_i^2\right] = \mathrm{Softmax}\big(\mathrm{MI}(\boldsymbol{z}_i^1, \boldsymbol{f}_i),\ \mathrm{MI}(\boldsymbol{z}_i^2, \boldsymbol{f}_i)\big) \tag{6}$$

where the softmax function is used to normalize mutual information scores into a probability distribution, ensuring positive contribution weights summing to unity. This enables intra-sample modality comparison and provides stable regularization.

Building upon this, to alleviate bias introduced by the dominant modality, the cross-entropy loss is combined with a regularization term that penalizes large disparities in modality contributions. This term promotes balanced feature extraction by encouraging the model to maintain fairness across modalities. The mathematical formulation is expressed as follows:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} |c_i^1 - c_i^2| \tag{7}$$

The total loss function integrates this regularization term with the classification loss $\mathcal{L}_{\text{cls}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{reg}} \tag{8}$$

By leveraging this mechanism, biases caused by modality imbalance can be effectively mitigated during the pretraining phase of the shared classifier.

3.3 Classifier-Constrained Alternating Training

Figure 3 illustrates our overall training pipeline. Following classifier pretraining, we perform modality-wise alternating training. A key challenge arises here: the classifier $\mathrm{Cls}(\cdot)$, which was adapted to the decision boundaries of the fused features $\boldsymbol{f}$ during pretraining, must now process unimodal features $\boldsymbol{z}^m$ during alternating training, where $P(\boldsymbol{z}^m|y) \neq P(\boldsymbol{f}|y)$. This distribution mismatch in the decision space disrupts feature continuity, leading to optimization instability and feature confusion Zhou et al. (2025a).

Algorithm 1 Classifier-Constrained Alternating Training (CCAT)

Inputs: multimodal dataset $\mathcal{D}$ with N samples and M modalities, epochs E, mini-batch $\mathcal{B} \sim \mathcal{D}$, threshold β
Parameters: pretrained shared classifier Cls, encoders $\{\mathrm{Enc}_m\}_{m=1}^{M}$, LoRA modules $\{\mathrm{LoRA}_m\}_{m=1}^{M}$
Outputs: trained encoders $\{\mathrm{Enc}_m\}_{m=1}^{M}$ and LoRA modules $\{\mathrm{LoRA}_m\}_{m=1}^{M}$
Freeze Cls; initialize $\{\mathrm{Enc}_m\}, \{\mathrm{LoRA}_m\}$
for epoch $e = 1, \ldots, E$ **do**
  for modality $m = 1, \ldots, M$ **do**
    Compute loss $\mathcal{L}_{\text{alt}}^{m}$ via Eq. (11) using modality-m data in $\mathcal{B}$
    Update Enc_m and LoRA_m via gradient descent
  end for
  Estimate contributions $\{c_i^1, \ldots, c_i^M\}_{i=1}^{|\mathcal{B}|}$ via Eq. (6)
  for modality $m = 1, \ldots, M$ **do**
    Construct subset $\mathcal{B}_m^{\text{extreme}} = \{\boldsymbol{x}_i^m \in \mathcal{B} \mid c_i^m < \beta\}$
    Compute loss $\mathcal{L}_{\text{retrain}}^{m}$ via Eq. (12) using modality-m data in $\mathcal{B}_m^{\text{extreme}}$
    Update Enc_m and LoRA_m via gradient descent
  end for
end for
return $\{\mathrm{Enc}_m\}_{m=1}^{M}, \{\mathrm{LoRA}_m\}_{m=1}^{M}$

To mitigate this issue, we integrate a lightweight Low-Rank Adaptation (LoRA) module for each modality Hu et al. (2022), as shown in Figure 3(c). Each LoRA module acts as a low-rank residual correction applied to the features:

$$\mathrm{LoRA}_m(\boldsymbol{z}_i^m) = \mathbf{B}^m \mathbf{A}^m \boldsymbol{z}_i^m \tag{9}$$

Specifically, in each iteration, given a batch of input samples $\boldsymbol{x}_i = [\boldsymbol{x}_i^1, \boldsymbol{x}_i^2]_{i=1,2,\ldots,B}$, the model sequentially processes the data of each individual modality across the entire batch using its corresponding modality-specific encoder and LoRA module, thereby obtaining the respective prediction outputs $\hat{y}_i^m$ and associated cross-entropy loss $\mathcal{L}_{\text{alt}}^m$:

$$\hat{y}_i^m = \mathrm{Softmax}\left(\mathrm{Cls}(\boldsymbol{z}_i^m) + \mathrm{LoRA}_m(\boldsymbol{z}_i^m)\right) \tag{10}$$

$$\mathcal{L}_{\text{alt}}^m = \frac{1}{B} \sum_{i=1}^{B} \mathrm{CE}(\hat{y}_i^m, y_i) \tag{11}$$

where B is the batch size; and y_i represents the ground-truth label of the i-th sample. Subsequently, the modality-specific cross-entropy loss $\mathcal{L}_{\text{alt}}^m$ is utilized to optimize the parameters of Enc_m through backpropagation.

To mitigate weak-modality under-optimization, we propose a sample-level secondary update. After initial full updates of all modality encoders and LoRAs, a modality contribution score c_i^m is computed for each sample $\boldsymbol{x}_i^m$, as defined in Equations (6) and (7), reflecting the contribution degree of modality m in predicting the target label. Notably, unlike the cross-attention fusion adopted in the first-stage training, here the computation of c follows the same decision-level fusion used in the inference stage, since this is also the method that produces the final prediction. We then identify highly imbalanced samples $\mathcal{B}_m^{\text{extreme}} = \{\boldsymbol{x}_i \in \mathcal{B} \mid c_i^m < \beta\}$ using threshold β. These samples are reprocessed through Enc_m and LoRA_m, and the frozen classifier to generate predictions. The auxiliary loss is defined as:

$$\mathcal{L}_{retrain}^m = \frac{1}{L} \sum_{i=1}^{L} \mathrm{CE}(\hat{y}_i^m, y_i), \tag{12}$$

where L denotes the number of samples in the re-training subset $\mathcal{B}_m^{\text{extreme}}$ and $\boldsymbol{y}_i$ is the ground-truth label of the i-th sample. This loss drives secondary gradient updates on Enc_m and LoRA_m, enhancing representation learning for modality-imbalanced instances. The whole trainingpipeline is provided in Algorithm 1.

During inference, each modality is processed by its dedicated encoder, transformed via corresponding LoRA modules, and classified through the shared classifier to generate unimodal predictions. These are fused at the decision level for final output.

Table 1: Performance comparison on CREMA-D, KS and MVSA with different baseline. Both the results of only using a single modality and the results of combining all modalities ("Multi") are listed. We report the average test accuracy (%) of three random seeds. The best results are highlighted in bold, and the second-best results are marked with a gray background.

Method	CREMA-D			Kinetic-Sound			MVSA		
	Multi	Audio	Video	Multi	Audio	Video	Multi	Image	Text
Sum	65.46	60.62	26.08	64.72	48.77	24.52	73.06	27.11	70.56
Concat	61.56	55.65	18.68	64.84	49.81	24.67	73.22	25.99	70.71
FiLM	60.07	53.89	18.67	63.33	48.67	23.15	75.34	27.12	74.85
BiGated	59.21	51.49	17.34	63.72	49.96	23.77	75.94	28.15	73.13
OGM-GE	68.14	53.76	28.09	65.78	51.57	32.19	76.37	31.98	74.76
QMF	63.71	59.41	39.11	65.78	29.73	32.19	77.96	32.99	74.87
MLA	80.78	63.17	68.01	71.35	54.67	51.03	75.14	53.37	73.22
MMPareto	75.13	65.46	55.24	70.13	56.40	53.05	78.81	**59.54**	74.76
LFM	83.62	63.17	45.83	72.53	54.12	**55.62**	-	-	-
CCAT (Ours)	**85.89**	**65.99**	**73.79**	**79.29**	**61.65**	53.75	**80.73**	55.30	**77.46**

Table 2: Ablation study (accuracy %) on components removed from the full method: Fix: classifier freezing (without LoRA); Alt: alternate training; Sec: secondary updates; LoRA: Low-Rank Adaption modules.

Fix	Alt	Sec	LoRA	CREMA-D			Kinetic-Sound			MVSA		
				Multi	Audio	Video	Multi	Audio	Video	Multi	Image	Text
✗	✓	✓	✓	82.80	64.38	71.77	77.26	59.78	**54.32**	78.03	53.95	77.07
✓	✗	✓	✓	81.45	64.92	69.62	77.47	**63.01**	50.68	78.32	54.91	74.76
✓	✓	✗	✓	83.06	65.59	69.49	78.25	61.97	52.03	79.38	54.62	75.34
✓	✓	✓	✗	84.68	65.19	73.25	78.77	62.64	53.01	80.35	55.17	76.49
✓	✓	✓	✓	**85.89**	**65.99**	**73.79**	**79.29**	61.65	53.75	**80.73**	**55.30**	**77.46**

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Datasets. To validate the effectiveness of the proposed method, we conduct experiments on three widely used multimodal datasets (see Appendix A.2 for detailed data descriptions): the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) Cao et al. (2014), featuring audio-visual recordings of American English speech acted with diverse emotional expressions; the Kinetic-Sound (KS) dataset Arandjelovic & Zisserman (2017), containing synchronized video-audio pairs for object and action recognition; and the Multimodal Visual Sentiment Analysis (MVSA) dataset Niu et al. (2016), focusing on sentiment classification in multimedia posts using both text and images. These datasets are representative in the context of modality imbalance, covering various modality combinations and reflecting different real-world imbalance patterns. This diverse selection ensures a comprehensive evaluation of our approach.

Baselines. We benchmark our method against conventional multimodal fusion approaches and recent state-of-the-art methods on the CREMA-D, KS and MVSA datasets. Baselines include: (i) Simple Fusion: Sum, Concat; (ii) Modulation-based Fusion: FiLM Perez et al. (2018), BiGated Kiela et al. (2018); (iii) Imbalance-aware Methods: OGM-GE Peng et al. (2022), QMF Zhang et al. (2023); and (iv) Recent SOTA: Multimodal Learning with Alternating Unimodal Adaptation (MLA) Zhang et al. (2024), MultiModal Pareto (MMPareto) Wei & Hu (2024), and LFM Yang et al. (2024).

Evaluation Metrics. We report multimodal accuracy (Acc) alongside unimodal performance for each baseline. When evaluating single modalities: (i) For FiLM, BiGated, OGM-GE and QMF, we disable the complementary modality within the fusion network; (ii) For Sum fusion Peng et al. (2022), features from the target modality are fed directly into the shared classifier head; (iii) For

Concat fusion Peng et al. (2022), we utilize dedicated subheads corresponding to each modality as per established protocol.

Implementation Details. We employ ResNet18 encoders for both audio and visual modalities across all datasets. For text-image data, image features are extracted with ResNet50 and textual features with BERT. All models were optimized via Stochastic Gradient Descent (SGD) with a batch size of 32, initial learning rate of 0.001, momentum of 0.9, and weight decay coefficient of 0.1. The learning rate was decayed by a factor of 10 every 70 epochs, with training conducted for 150 total epochs. The regularization coefficient λ for the loss function in classifier pre-training is set to 0.001. Experimental analysis reveals negligible performance sensitivity to LoRA's scaling factor α due to the classifier's limited parameter scale, leading to its universal fixation at $\alpha = 1$. The hyperparameter tuning was conducted in a sequential manner: first, the optimal LoRA rank r was selected from $1, 2, 4, 8, 16$, followed by the optimization of the modality imbalance threshold β from $0.05, 0.10, \ldots, 0.40$, via validation-set grid searches. As detailed in Table 3 and Figure 4, configurations are empirically set to $(r = 2, \beta = 0.15)$ for CREMA-D, $(r = 2, \beta = 0.30)$ for KS, and $(r = 8, \beta = 0.05)$ for MVSA to identify severely imbalanced samples. All experiments execute on NVIDIA RTX 4090 GPUs. Additional experimental details are provided in the Appendix A.3.

Table 3: Grid search results for LoRA rank r.

Dataset	LoRA Rank r				
	1	2	4	8	16
CREMA-D	84.41	**85.35**	84.68	84.81	84.54
KS	78.46	**78.83**	78.41	78.67	78.76
MVSA	76.49	79.00	74.57	**79.58**	79.38

4.2 Comparison with SOTA MML Baselines

Table 1 summarizes main results across all datasets, revealing three key observations: (i) Our method substantially outperforms all baselines, including conventional multimodal learning and modality balancing techniques, achieving state-of-the-art performance in most scenarios; (ii) All modality rebalancing techniques significantly surpass traditional feature concatenation or summation, confirming both the performance penalty induced by modality imbalance and the efficacy of balancing strategies; (iii) Unlike prior works Zhang et al. (2024); Yang et al. (2024) equating reduced unimodal gaps with balance, we prioritize liberating weak modalities representational potential. Their significant accuracy gains across benchmarks directly validate imbalance mitigation transcending relative performance differences; (iv) For MLA, MMPareto, LFM, and our method CCAT, unimodal results are directly acquired from decision-level fusion outputs.

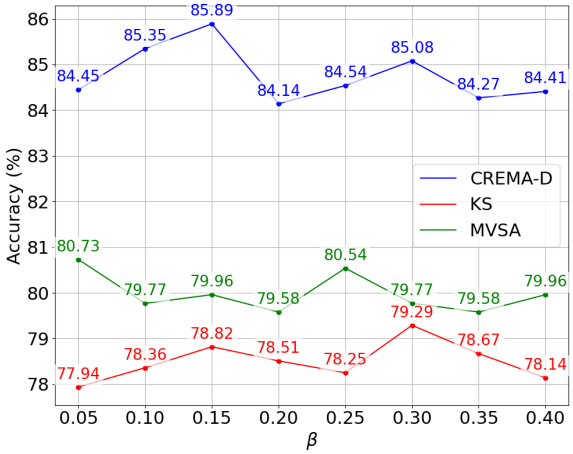
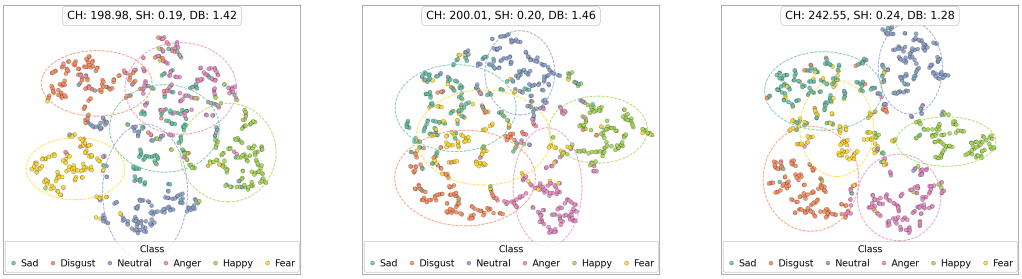


Figure 4: Grid search results for modality imbalance thresholds β on the validation set. The optimal combination is selected based on the highest validation accuracy.

4.3 Ablation Study

Table 2 presents ablation results on the CREMA-D dataset (full results in Appendix). Our study systematically validates CCAT's efficacy in mitigating modality imbalance: (i) Classifier freezing significantly enhances suppressed modalities during alternating training, mitigating inter-modal gradient conflicts while blocking dominant modalities from monopolizing decision boundaries; (ii) Secondary updates deliver targeted enhancement to underperforming modalities by suppressing overconfidence in dominant modalities, thereby calibrating cross-modal dynamics at the sample level;



(a) MLA Model (b) Non-Fixed Model (c) CCAT Model (Ours)

Figure 5: Visualizations of feature distributions based on t-SNE for test samples under MLA (a), Non-Fixed Classifier (b), and Our Proposed Method (c). Calinski-Harabasz (CH), Silhouette (SH), and Davies-Bouldin (DB) scores are computed to quantitatively assess the clustering quality. Compared to the other methods, our approach achieves clearer class separability, especially improving the distinction of the fear and sad classes from other categories.

(iii) LoRA adapters substantially boost multimodal fusion performance while preserving modality-specific characteristics, confirming their capacity to orchestrate shared classification knowledge and modality-exclusive features. The integrated framework achieves optimal performance, demonstrating CCAT's effectiveness in mitigating modality imbalance.

4.4 Further Analysis

Enhancing Discriminative Space via Fixed Classifier Design. To investigate whether freezing the classifier contributes to constructing more discriminative decision boundaries, we visualized the distribution of all test samples in the feature space using t-SNE projections, as shown in Figure. 5. Compared with the MLA baseline and the non-fixed classifier setting, our proposed method exhibits improved class separability, especially for the *fear* and *sad* classes, which become more clearly separated from other categories.

Beyond qualitative visualization, we perform quantitative clustering analysis using standard metrics: Calinski-Harabasz (CH) Caliński & Harabasz (1974), Silhouette (SH) Rousseeuw (1987), and Davies-Bouldin (DB) Davies & Bouldin (1979). As shown in Figure. 5, our method achieves optimal clustering quality, highest CH and SH scores and lowest DB score, demonstrating superior intra-class compactness and inter-class separation. These results confirm that the fixed-classifier strategy yields more discriminative feature representations.

5 Conclusion

This work addresses modality imbalance through a classifier-centric paradigm. Inspired by class imbalance remedies, we propose Classifier-Constrained Alternating Training (CCAT) that bridges both problems via optimization dynamics analysis. Our two-stage framework systematically mitigates imbalance: (1) Pretraining a shared classifier with contribution-aware regularization yields unbiased initialization; (2) Freezing this classifier during modality-alternating optimization provides stable objectives, while lightweight LoRA adapters enable modality-specific adaptation. Complementary sample-level re-optimization further enhances underrepresented modalities. CCAT consistently outperforms existing methods across benchmarks, validating its efficacy for discriminative representation learning in imbalanced multimodal scenarios.

6 Future Work

In the future, we plan to extend the proposed framework to more complex scenarios involving tri-modal datasets. A key direction is to investigate the effectiveness of our modality-wise alternating training and imbalance-aware fine-tuning strategy when three distinct modalities are present.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

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, 2019. doi: 10.1109/TPAMI.2018.2798607.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

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.

Kailei Cheng, Lihua Tian, and Chen Li. Lawnet: Audio-visual emotion recognition by listening and watching. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

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*, pp. 8632–8656. PMLR, 2023.

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*, pp. 20029–20038, 2023.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21630–21643. Curran Associates, Inc., 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*, 2024.

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

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.

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*, pp. 22214–22224, 2023.

Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. Af: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9236–9254, 2021.

Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 1039–1047, 2023.

Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Calibrating multimodal learning. In *International Conference on Machine Learning*, pp. 23429–23450. PMLR, 2023.

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

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 (CVPR)*, pp. 8238–8247, June 2022.

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.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Sitong Su, Junchen Zhu, Lianli Gao, and Jingkuan Song. Utilizing greedy nature for multimodal conditional image synthesis in transformers. *IEEE Transactions on Multimedia*, 26:2354–2366, 2023.

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.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

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 (CVPR)*, June 2020.

Yake Wei and Di Hu. Mmpareto: boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024.

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*, pp. 27338–27347, 2024a.

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*, 2024b.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 17–23 Jul 2022.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Shaoxuan Xu, Menglu Cui, Chengxiang Huang, Hongfa Wang, and Di Hu. Balancebenchmark: A survey for multimodal imbalance learning, 2025. URL https://arxiv.org/abs/2502.10816.

Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu. Domfn: A divergence-orientated multi-modal fusion network for resume assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1612–1620, 2022a.

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.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4553–4566, 2025. doi: 10.1109/TPAMI.2025.3547417.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35:37991–38002, 2022b.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

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*, pp. 27456–27466, 2024.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20852–20862, 2025a.

Ying Zhou, Xuefeng Liang, Yue Xu, and Bowen Gao. Sample-level self-paced learning to tackle multimodal imbalance problem. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

A Appendix

A.1 Implementation of Bidirectional Cross-Attention Mechanism

Consider a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1,2,\ldots,N}$, where each sample $\boldsymbol{x}_i = [\boldsymbol{x}_i^1, \boldsymbol{x}_i^2]$ contains the audio modality $\boldsymbol{x}_i^1$ and the visual modality $\boldsymbol{x}_i^2$, accompanied by its ground-truth label y_i. Modality-specific encoders $(\mathrm{Enc}_1, \mathrm{Enc}_2)$ are employed to extract features from each modality independently:

$$\boldsymbol{z}_i^1 = \mathrm{Enc}_1(\boldsymbol{x}_i^1), \quad \boldsymbol{z}_i^2 = \mathrm{Enc}_2(\boldsymbol{x}_i^2) \tag{13}$$

To comprehensively model the interactions between modalities, we adopt a bidirectional cross-attention (Bi-Cross Attention) mechanism for feature fusionCheng et al. (2024). This mechanism consists of cross-attention calculations in two directions: from vision to audio and from audio to vision, formulated as follows:

Visual-to-Audio Fusion. In this direction, the audio features $\boldsymbol{z}_i^1$ are treated as the Query, while the visual features $\boldsymbol{z}_i^2$ serve as both the Key and the Value. The multi-head attention (MHA) output $\boldsymbol{a}_i^1$ is first computed as:

$$\boldsymbol{a}_i^1 = \mathrm{MHA}(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2, \boldsymbol{z}_i^2) \tag{14}$$

To stabilize and enhance the feature fusion representation, a residual connection is then added, followed by layer normalization $(\mathrm{LayerNorm})$:

$$\boldsymbol{h}_i^1 = \mathrm{LayerNorm}\big(\boldsymbol{z}_i^1 + \boldsymbol{a}_i^1\big) \tag{15}$$

Subsequently, a feed-forward network (FFN) is applied to introduce non-linearity and model higher-level interactions, with another layer normalization $(\mathrm{LayerNorm})$ step performed thereafter:

$$\boldsymbol{f}_i^1 = \mathrm{LayerNorm}\big(\boldsymbol{h}_i^1 + \mathrm{FFN}(\boldsymbol{h}_i^1)\big) \tag{16}$$

Finally, average pooling $(\mathrm{AvgPool})$ is employed to aggregate the sequence into a compact fused representation in the visual-to-audio direction:

$$\boldsymbol{f}_i^{2\to1} = \mathrm{AvgPool}(\boldsymbol{f}_i^1) \tag{17}$$

Audio-to-Visual Fusion. In this direction, the visual features $\boldsymbol{z}_i^2$ are designated as the Query, while the audio features $\boldsymbol{z}_i^1$ serve as the Key and Value. The computation proceeds analogously as follows:

$$\boldsymbol{a}_i^2 = \mathrm{MHA}(\boldsymbol{z}_i^2, \boldsymbol{z}_i^1, \boldsymbol{z}_i^1) \tag{18}$$

$$\boldsymbol{h}_i^2 = \mathrm{LayerNorm}(\boldsymbol{z}_i^1 + \boldsymbol{a}_i^2) \tag{19}$$

$$\boldsymbol{f}_i^2 = \mathrm{LayerNorm}\big(\boldsymbol{h}_i^2 + \mathrm{FFN}(\boldsymbol{h}_i^2)\big) \tag{20}$$

$$\boldsymbol{f}_i^{1\to 2} = \mathrm{AvgPool}(\boldsymbol{f}_i^2) \tag{21}$$

Subsequently, the bidirectionally fused feature representation $\boldsymbol{f}_i$ is obtained by summing the fused features derived from both the audio-to-visual and visual-to-audio directions:

$$\boldsymbol{f}_i = \boldsymbol{f}_i^{1\to 2} + \boldsymbol{f}_i^{2\to 1} \tag{22}$$

This unified representation is then passed through a shared classifier $cls(\cdot)$ to generate the final prediction:

$$\hat{y}_i = cls(\boldsymbol{f}_i) \tag{23}$$

A.2 Detailed Dataset Description

A.2.1 CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)

CREMA-D is a widely used multimodal benchmark for emotion recognition, comprising recordings from 91 actors of diverse demographic backgrounds. Each actor performed 12 scripted sentences with six categorical emotions (anger, disgust, fear, happiness, neutral, sadness) expressed at varying intensity levels. The dataset was annotated through large-scale crowdsourcing, enabling rigorous investigation of unimodal versus multimodal affect perception and cross-demographic variations in emotional expression.

A.3 KS (Kinetic-Sound)

Kinetic-Sound is an audiovisual benchmark dataset derived from the Kinetics corpus for multimodal action recognition research. It depict 31 visually and aurally distinctive real-world actions. Curated specifically for multimodal learning, the dataset supports investigations into audio-visual correspondence, self-supervised representation learning, and complementary information fusion. Its YouTube-sourced content facilitates robust model development for dynamic scene understanding where both visual motion and auditory signatures contribute to action classification.

A.4 MVSA (Multimodal Visual Sentiment Analysis)

MVSA serves as a benchmark for multimodal sentiment analysis in social media contexts, containing Twitter-derived images paired with textual captions and sentiment polarity labels (positive/neutral/negative). This authentic user-generated content captures complex cross-modal sentiment interactions, supporting research on multimodal fusion, cross-modal alignment, and sentiment disambiguation where textual and visual cues exhibit amplification or conflict.

A.5 Details of experimental setup

A.5.1 Encoders

To evaluate the effectiveness of our method across diverse models and datasets, we employed three distinct encoders:

ResNet-18. We employed ResNet-18 as the visual-encoder backbone for the CREMA-D and KS datasets. ResNet-18 belongs to the pioneering Residual Network family introduced in 2015, which overcame longstanding training challenges in deep neural architectures through the use of residual connections. This design facilitates direct gradient flow and significantly mitigates the vanishing-gradient and degradation issues that afflict plain deep networks. We initialized ResNet-18 with the standard weight initialization strategy commonly prescribed in deep learning frameworks.

ResNet-50. For the image modality of the MVSA dataset, we employ ResNet50, a deeper variant within the Residual Network (ResNet) family. Like its shallower counterparts, each block includes an identity shortcut that bypasses nonlinear transformations, enabling the model to learn only the residual mapping and thereby ensuring effective training of deeper networks. We initialized the weights of ResNet-50 by standard initialization.

BERT. Our textual encoder is based on BERT (Bidirectional Encoder Representations from Transformers). This model consists of 12 Transformer layers, each with 768 hidden dimensions and 12 attention heads, totaling approximately 110 million parameters. We employed the official BERT-Base pretrained checkpoint, as released by the authors, to initialize our encoder. This checkpoint has become a standard foundation for downstream fine-tuning across diverse NLP tasks. In our setup, the pretrained BERT-Base model served as a feature encoder, upon which task-specific layers are optionally stacked and optimized.

A.6 Implementation Details

A unified classification head comprising a single fully connected layer was implemented across all experimental configurations. The input dimensionality was configured at 512 for base model experiments and 768 for large pretrained model evaluations. Audio processing pipelines resampled raw waveforms to 22,050 Hz, extending clips to a fixed 20-second duration through signal replication. Log-magnitude spectrograms were generated via short-time Fourier transform (STFT) using 512-point FFT windows with 353-sample hop lengths. For visual inputs, three frames were uniformly sampled per video clip and resized to 224×224 resolution, with standard augmentation techniques and ImageNet normalization applied during training. Textual features were derived from BERT-based tokenized representations.

A.7 Large Language Models Statement of Use

We acknowledge the use of Large Language Models (LLMs) exclusively for text refinement, including grammar correction, style polishing, and improving readability. No LLMs were involved in designing the methodology, conducting experiments, analyzing results, or drawing scientific conclusions. All technical contributions, experimental implementations, and analyses were performed solely by the authors.