ReconBoost: Boosting Can Achieve Modality Reconcilement

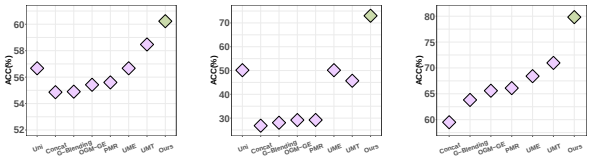
Cong Hua [1 2] **Qianqian Xu** [1] **Shilong Bao** [3 4] **Zhiyong Yang** [2] **Qingming Huang** [2 1 5]

Abstract

This paper explores a novel multi-modal *alternating* learning paradigm pursuing a reconciliation between the exploitation of uni-modal features and the exploration of cross-modal interactions. This is motivated by the fact that current paradigms of multi-modal learning tend to explore multi-modal features simultaneously. The resulting gradient prohibits further exploitation of the features in the weak modality, leading to modality competition, where the dominant modality overpowers the learning process. To address this issue, we study the modality-alternating learning paradigm to achieve reconcilement. Specifically, we propose a new method called *ReconBoost* to update a fixed modality each time. Herein, the learning objective is dynamically adjusted with a reconcilement regularization against competition with the historical models. By choosing a KL-based reconcilement, we show that the proposed method resembles Friedman's Gradient-Boosting (GB) algorithm, where the updated learner can correct errors made by others and help enhance the overall performance. The major difference with the classic GB is that we only preserve the newest model for each modality to avoid overfitting caused by ensembling strong learners. Furthermore, we propose a memory consolidation scheme and a global rectification scheme to make this strategy more effective. Experiments over six multi-modal benchmarks speak to the efficacy of the method. We release the code at https://github.com/huacong/ReconBoost.

[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China [2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China [3]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China [4]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China [5]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China. Correspondence to: Qianqian Xu <xuqianqian@ict.ac.cn>, Qingming Huang <qmhuang@ucas.ac.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) Audio Modality (b) Visual Modality (c) Multi-modal

Figure 1. The performance among multi-modal learning competitors on the CREMA-D dataset. For audio modality and visual modality, we evaluate the encoders of different competitors by training linear classifiers on them. Uni represents the uni-modal training method.

1. Introduction

Deep learning has significantly advanced uni-modal tasks (He et al., 2016; Tang et al., 2017; Li et al., 2019; Zhang et al., 2020). However, most real-world data usually follows a multi-modal nature (say text, video, and audio) in various fields such as data mining (Cai et al., 2011; Jiang et al., 2019), computer vision (Gallego et al., 2022; Wan et al., 2023; Shao et al., 2024), and medical diagnosis (Chen et al., 2022; Ruan et al., 2021). Because of this, the deep learning community has recently focused more on multi-modal learning (Wei et al., 2022; Jiang et al., 2023a; Feng et al., 2022; Zhang et al., 2024). The prevailing paradigm in multi-modal learning typically employs a **joint learning** strategy, wherein a wealth of studies (Shahroudy et al., 2017; Chen et al., 2020; Wang et al., 2020b; Deng & Dragotti, 2021; Zhang et al., 2023; Jiang et al., 2023b; Shao et al., 2023) primarily focus on integrating modality-specific features into a **shared representation** for various downstream tasks.

Despite great success, numerous experimental observations (Du et al., 2023; Peng et al., 2022) and recent theoretical evidence (Huang et al., 2022) have pointed out that current paradigms of multi-modal learning encounter ***Modality Competition,*** where the model is dominated by some of the modalities. Various studies (Peng et al., 2022; Fan et al., 2023; Du et al., 2023) have been made to mitigate the modality competition issue. The primary concern is how to balance optimization progress across multi-modal learners and improve uni-modal feature exploitations. For instance, G-Blending (Wang et al., 2020a) adds a uni-modal classifier with additional supervised signals in multi-modal

learning to blend modalities effectively. OGM (Peng et al., 2022) and PMR (Fan et al., 2023) work on reducing the influence of the dominant modality and aiding the training of others through adaptive gradient modulation and dynamic loss functions, respectively. Besides, UMT (Du et al., 2023) distills knowledge from well-trained uni-modal models in multi-modal learning which can effectively benefit from uni-modal learning.

However, as depicted in Fig.1, existing algorithms **still follow the joint learning strategy**, suffering from limited performance trade-offs for modality competition. Expanding the gradient update rule, we find that joint learning tends to neglect the gradient from weak modality. The dominant modality that converges more quickly would eventually overpower the whole learning process.

Therefore, in this paper, we turn to the following question:

Can we achieve modality reconciliation via other learning paradigms?

In search of an answer, we propose an effective method named ***ReconBoost***, where we alternate the learning for each modality. Intuitively, it naturally alleviates modality competition in the gradient space since the modality-specific gradients must be employed separately. To further enhance the effect of individual modalities, we propose a reconcilement regularization to maximize the diversity between the current update and historical models. Dynamically adjusting the learning objective via the regularization term further alleviates the modality competition issue induced by sticking to a particular modality. Theoretically, we show that by choosing a KL divergence (Kullback & Leibler, 1951) based reconcilement term, our proposed method can realize an alternating version of the well-known gradient boosting method (Friedman, 2001). Specifically, the updated modality learner can focus on the errors made by others, thereby highlighting their complementarity. Unlike traditional boosting techniques (Freund, 1995; Freund & Schapire, 1997), which use weak learners like decision trees, our method employs DNN-based learners which are over-parameterized models. To avoid overfitting, we discard historical learners and only preserve the last learner for each modality, creating an ***alternating-boosting strategy.*** Additionally, considering the differences between the traditional boosting techniques and our alternating-boosting strategy, we present a memory consolidation scheme and a global rectification scheme to reduce the risk of forgetting critical patterns in historical updates.

Finally, we conduct empirical experiments on six multi-modal benchmarks and demonstrate that 1) ReconBoost can consistently outperform all the competitors significantly on all datasets. 2) ReconBoost can achieve ***Modality Reconcilement.***

2. Preliminary

In this section, we first review the task of multi-modal learning. Then, we further explain the current difficulties encountered in multi-modal learning. ***Due to space limitations, we present a brief overview of prior arts in App.A.***

2.1. The Task of Multi-modal Learning

Let $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ be a multi-modal dataset, where N is the number of examples in the training set. Herein, each example i consists of a set of raw features $x_i = \{m_i^k\}_{k=1}^{M}$ from different M modalities and a one-hot label $y_i = \{c_{i,j}\}_{j=1}^{Y}$, where $c_{i,j} = 1$ if the label for i is j, otherwise $c_{i,j} = 0$; Y is the total number of categories.

Given M modality-specific feature extractors $\{\mathcal{F}_k(\theta_k)\}_{k=1}^{M}$, with $\mathcal{F}_k$ typically being a deep neural network with parameters θ_k, $\{\mathcal{F}_k(\theta_k; m_i^k)\}_{k=1}^{M}$ denotes the latent features of i-th sample, where $\mathcal{F}_k(\theta_k; m_i^k) \in \mathbb{R}^{d_k}$. Then, we define the predictor $\mathcal{S}$ as a mapping from the latent feature space to the label space. The **objective** of multi-modal learning is to jointly minimize the empirical loss of the predictor:

$$\mathcal{L}(\mathcal{S}(\{\mathcal{F}_k(x)\}_{k=1}^{M}), y) = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathcal{S}(\{\mathcal{F}_k(\theta_k; m_i^k)\}_{k=1}^{M}), y_i) \tag{1}$$

where ℓ is the CE loss. In multi-modal learning, a key step is to merge the modality-specific representations. To this end, the predictor is often formalized as a composition: $g \circ f$, where g is a simple classifier and f is a cross-modal fusion strategy. For example, one can use the concatenate operation to implement the fusion strategy and use a linear model to implement the classifier. In this case, the resulting predictor becomes:

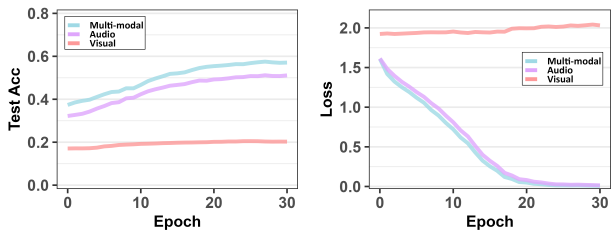
$$\begin{aligned}\mathcal{S}(\{\mathcal{F}_k(\theta_k; m_i^k)\}_{k=1}^{M}) &= W \cdot [\mathcal{F}_1(\theta_1; m_i^1) : \cdots : \mathcal{F}_M(\theta_k; m_i^k)] \\ &= \sum_k W_k \cdot \mathcal{F}_k(\theta_k; m_i^k),\end{aligned} \tag{2}$$

where $W \in \mathbb{R}^{Y \times \sum_k d_k}$ is the last linear classifier, $W_k \in \mathbb{R}^{Y \times d_k}$ is a block of W for the k-th modality.

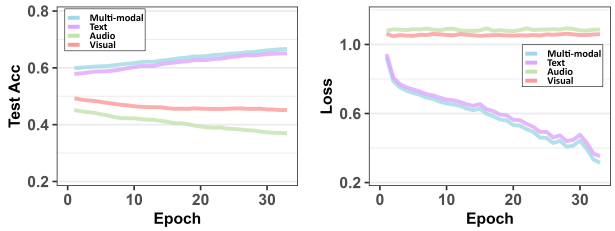
In contrast to uni-modal training, information fusion in multi-modal learning can help explore cross-modal interactions, enhancing performance across various real-world scenarios. However, under the current paradigm of multi-modal learning, the limitations in effectively exploiting uni-modal features have constrained the performance of the multi-modal learning model. We state the corresponding challenges herein in the upcoming subsection.

2.2. The Challenge of Multi-modal Learning

Current paradigms ***synchronously*** optimize the objective across all modalities. In this setting, a joint gradient descent



(a) Testing Acc. on CREMA-D (b) Training Loss on CREMA-D



(c) Testing Acc. on MOSEI (d) Training Loss on MOSEI

Figure 2. The phenomenon of modality competition is observed in the concatenation fusion method when applied to two datasets: CREMA-D with two modalities and MOSEI with three modalities. In CREMA-D, the learning process is primarily influenced by the audio modality, leading to insufficient learning of the visual modality. In MOSEI, the text modality takes control of multi-modal learning, causing challenges in updating the parameters of both the audio and visual modalities.

update will trigger the modality competition. To see this, we expand the update rule for each modality. Here, we denote the merged score function as:

$$\Phi_M^t(x_i) = \sum_{k=1}^{M} W_k^t \cdot \mathcal{F}_k(\theta_k^t; m_i^k) \tag{3}$$

Then the update for the k-th modality-specific parameters can be written as:

$$\begin{aligned}\theta_k^{t+1} &= \theta_k^t - \eta \cdot \nabla_{\theta_k^t} \mathcal{L}(\Phi_M^t(x), y) \\ &= \theta_k^t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial W_k \cdot \mathcal{F}_k(\theta_k^t; m_i^k)}{\partial \theta_k^t} \right)^\top \\ &\qquad \cdot \underbrace{\frac{\partial \ell\left(\Phi_M^t(x_i), y_i\right)}{\partial \Phi_M^t(x_i)}}_{\text{shared}}\end{aligned}$$

$$\begin{aligned}W_k^{t+1} &= W_k^t - \eta \cdot \nabla_{W_k^t} \mathcal{L}(\Phi_M^t(x), y) \\ &= W_k^t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{F}_k(\theta_k^t; m_i^k) \right)^\top \cdot \underbrace{\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)}}_{\text{shared}},\end{aligned}$$

where η is the learning rate. The modality competition issue arises from the shared score gradient across modalities:

$$\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)} = \sigma_i^t - y_i \tag{4}$$

where $\sigma_i \in \mathbb{R}^Y$ means the prediction score for the i-th example. Once the gradient for the shared score is small, the update for all modalities will be stuck simultaneously. A modality k is said to have a consistent gradient with the shared scoring function if $\frac{\partial \ell(\phi_k^t(x_i), y_i)}{\partial \phi_k^t(x_i)}$ strongly resembles $\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)}$, where ϕ_k is the uni-modal score for modality k. If not, we consider modality k to have an inconsistent gradient with the shared scoring function. If the modality k has a consistent gradient with the shared score, then we can learn it well under this setting. On the opposite, if modality k has an inconsistent gradient with the shared score, it will be stuck at bad local optimums, leading to performance degradation. This phenomenon is depicted in Fig.2. To address this issue, our goal is to achieve reconcilement among modalities, where one can find a better trade-off between the exploitation of modality-specific patterns and the exploration of modality-invariance patterns.

Despite recent efforts to design various strategies (S mentioned above) in multi-modal learning to avoid modality competition, only limited improvements can be achieved, given the nature of synchronous optimization. The limitations inspire us to explore a modality-alternating learning strategy.

3. Methodology

In this section, we propose a modality-alternating framework called ReconBoost. The overall framework is illustrated in Fig.3. On top of the alternating update rule, we also incorporate a reconcilement regularization strategy to maximize the diversity between the current and historical models. Further details on ReconBoost will be discussed in the following.

3.1. Modality-alternating Update with Dynamic Reconcilment

Notations. Given M modality-specific classifiers $\{W_k\}_{k=1}^M$, along with modality-specific feature extractors, $\{\phi_k(\vartheta_k)\}_{k=1}^M$ represents M modality learners, where $\phi_k(\vartheta_k) = W_k \cdot \mathcal{F}_k(\theta_k)$. ϑ_k represents the parameters of the k-th modality learner and $\phi_k(\vartheta_k; m_i^k) \in \mathbb{R}^Y$.

We first introduce the naive version of modality-alternating learning.

Step 1: Each time, we pick a specific modality learner ϕ_k to update, and keep the others fixed. The gradient rule is formalized as follows:

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \mathcal{L}\left(\phi_k^t(m^k), y\right) \tag{5}$$

where

$$\mathcal{L}\left(\phi_k^t(m^k), y\right) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(\phi_k(\vartheta_k^t; m_i^k), y_i\right) \tag{6}$$

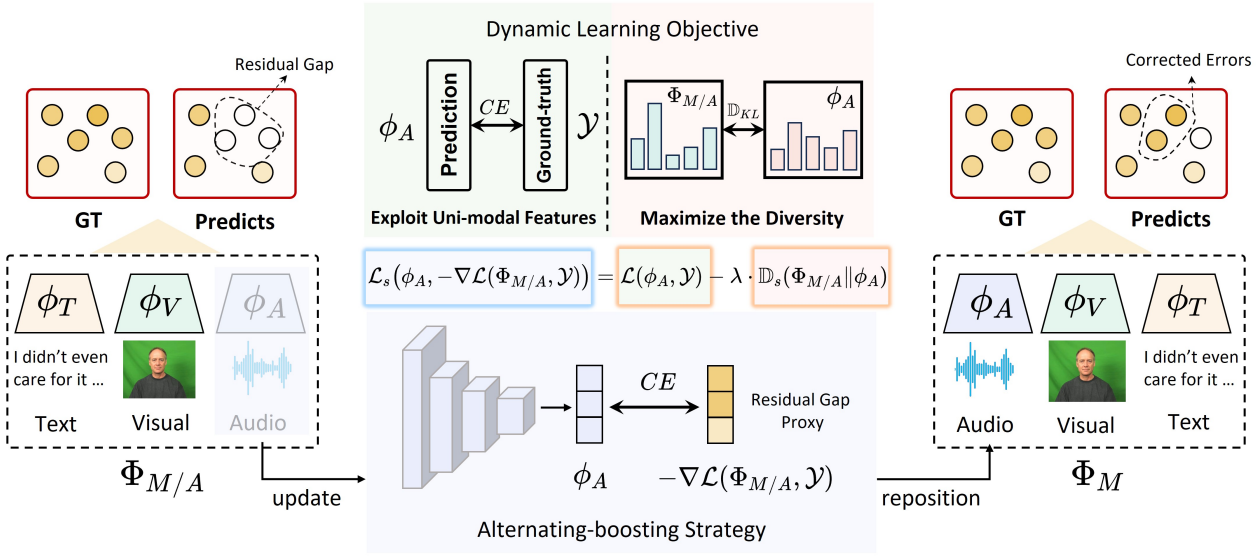


Figure 3. The overview of proposed ReconBoost. In round s, we pick up a specific modality learner to update and keep the others fixed. The updated modality learner can correct errors and enhance the overall performance.

is the loss for the k-th modality.

Step 2: After the alternating training procedure, multi-modal features are merged in the following way to produce the final score $\Phi_M(x_i) = \sum_{k=1}^{M} \phi_k(\vartheta_k; m_i^k)$.

When model updates are alternated, the gradients across different modalities are naturally disentangled from each other, alleviating the modality competition issue. While this approach ensures the exploitation of uni-modal features, it neglects the investigation of cross-modal diversity, limiting overall performance. It motivates us to design a more effective modality-specific supervised signal.

At each fixed time s in the update in step 1), we explore reconcilement regularization by introducing the following term in the loss:

$$\mathbb{D}_s\left(\Phi_{M/k}(x_i), \phi_k(\vartheta_k; m_i^k)\right)$$

where $\Phi_{M/k}(x_i) = \sum_{j=1, j\neq k}^{M} \phi_j(\vartheta_j; m_i^j)$. Here, $\mathbb{D}_s$ could be regarded as a diversity measure between the current block being updated and the historical models in the updating sequence. In pursuit of a dynamical reconcilement, in round s in Step 1), we turn to use the following objective:

$$\tilde{\mathcal{L}}_s(\phi_k(m^k), y) = \frac{1}{N}\sum_{i=1}^{N}\Bigg[\underbrace{\ell\left(\phi_k(\vartheta_k; m_i^k), y_i\right)}_{\text{agreement term}} - \lambda \cdot \underbrace{\mathbb{D}_s\left(\Phi_{M/k}(x_i), \phi_k(\vartheta_k; m_i^k)\right)}_{\text{reconcilement regularization term}}\Bigg] \tag{7}$$

In this new formulation, the loss function is no longer the same. In each round, we try to dynamically maintain the trade-off between **the agreement item** to align the overall predictor with the ground truth and **the reconcilement regularization term** to leverage complementary information between modalities. The parameter λ is a trade-off coefficient. The exploration of the impact of λ on the performance is presented in Sec.4.4 ablation experiments.

3.2. Connection to the Boosting Strategy: Theoretical Guanratee

At first glance, the dynamical variation of the loss function makes the optimization property of ReconBoost unclear. To further explore its theoretical foundation, we investigate the connection with the well-known Gradient-Boosting (GB) method (Freund, 1995; Friedman, 2001; Freund et al., 1996; Freund & Schapire, 1997), which is a powerful boosting method for additive expansion of models. The theoretical result is shown as follows:

Theorem 3.1. *When the reconcilement regularization satisfies,*

$$\lambda \cdot \nabla_{\phi_k} \mathbb{D}_s\left(\Phi_{M/k}(x_i), \phi_k(\vartheta_k; m_i^k)\right) = \nabla_{\phi_k} \ell\left(\phi_k(\vartheta_k; m_i^k), y_i\right) - \nabla_{\phi_k} \ell\left(\phi_k(\vartheta_k; m_i^k), -\nabla_{\Phi_{M/k}} \ell(\Phi_{M/k}(x_i), y_i)\right)$$

It leads to equivalent optimization goals:

$$\nabla_{\vartheta_k} \tilde{\mathcal{L}}_s\left(\phi_k(m^k), y\right) \iff \nabla_{\vartheta_k} \mathcal{L}\big(\phi_k(m^k), -\nabla_{\Phi_{M/k}} \ell(\Phi_{M/k}(x), y)\big)$$

Please refer to App.B for the proof in detail.

Here, to better understand the generality of our method and theorem, we will consider the case where the optimization loss function is CE loss. As a corollary of the theorem we have:

Corollary 3.2. *Let the reconcilement regularization be a KL divergence (Kullback & Leibler, 1951) function:*

$$\mathbb{D}_s\left(\Phi_{M/k}(x_i), \phi_k(\vartheta_k; m_i^k)\right) = \mathbb{D}_{KL,s}\left(\Phi_{M/k}(x_i) \| \phi_k(\vartheta_k; m_i^k)\right)$$

Then,

$$\nabla_{\vartheta_k} \tilde{\mathcal{L}}_s\left(\phi_k(m^k), y\right) \iff \nabla_{\vartheta_k} \mathcal{L}\big(\phi_k(m^k), -\nabla_{\Phi_{M/k}} \ell(\Phi_{M/k}(x), y)\big)$$

where ℓ is the CE loss.

Similar to the GB algorithm, optimizing the dynamic loss function $\tilde{\mathcal{L}}$ in ReconBoost consistently optimizes the original loss $\mathcal{L}$ with a progressively changing pseudo-label $-\nabla_{\Phi_{M/k}} \ell(\Phi_{M/k}(x), y)$. The pseudo-label is a gradient descent step at the space of Φ for the current time. The major difference from traditional GB is that we only employ the sum of the last updates for each modality, creating an ***alternating-boosting strategy.*** This is a **selective** additive expansion of the gradient decent on the functional space. This could be considered an **implicit bias** when the weak learners in traditional GB are replaced with over-parametrized deep learning models.

3.3. Enhancement Schemes

In this subsection, we elaborate on two enhancement schemes in ReconBoost, memory consolidation regularization, and global rectification scheme.

Memory Consolidation Regularization (MCR)**.** In contrast to GB, our alternating-boosting strategy preserves the newest learner for each modality while forgetting historical ones. Each updated modality learner fits the residual and effectively corrects errors from others. However, forgetting may result in modality-specific learners struggling with samples where others excel. To compensate for the discards, we propose MCR to enhance the performance of the modality-specific learner, formalized as:

$$\begin{aligned}&\mathcal{L}_{mcr}(-\nabla_{\phi_{k-1}} \ell(\phi_{k-1}(m^{k-1}), y), -\nabla_{\phi_k} \ell(\phi_k(m^k), y)) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla_{\phi_k} \ell(\phi_k(m_i^k), y_i) - \nabla_{\phi_{k-1}} \ell(\phi_{k-1}(m_i^{k-1}), y_i) \right\|^2\end{aligned} \tag{8}$$

where ϕ_{k-1} represents the previous modality learner. Intuitively, optimizing Eq.8 ensures that the predictions of ϕ_k will not be too far from that of ϕ_{k-1}, avoiding excessive focus on errors and benefiting consolidating memory of the modality-specific learner.

Global Rectification Scheme (GRS)**.** Following the standard paradigm of boosting, only the parameters of the k-th weak learner get updated during step k to greedily fit the residual, leaving the parameters of other learners unchanged. However, when dealing with modality learners implemented as over-parameterized neural networks in our alternating-boosting strategy, greedy learning strategy in the standard paradigm of boosting may cause the ensemble multi-modal learning model to fall in poor local minima easily, hindering the optimization of the objective. Drawing inspiration from (Badirli et al., 2020), we introduce GRS to overcome the challenge. After each update of the modality learner, instead of keeping the parameters of the $k-1$ modality learners fixed, we allow their parameters to be updated:

$$\vartheta_m^t = \vartheta_m^{t-1} - \eta \nabla_{\vartheta_m^{t-1}} \mathcal{L}(\Phi_M^{t-1}(x), y), \forall m \in [1, M] \tag{9}$$

where Φ_M represents adding the updated ϕ_k to $\Phi_{M/k}$; t means the t-th iteration in the rectification stage and η is the learning rate. In summary, these two schemes will enhance the performance of the proposed alternating-boosting strategy. Moreover, they provide insights into applying boosting techniques in the deep learning community.

3.4. Final Goal

Upon completing a cycle involving M stages, we reach the overall optimization objective for our proposed method in Eq.10.

$$\begin{aligned}\mathcal{L}_{all} = &\underbrace{\sum_{k\in[1,M]} \mathcal{L}(\phi_k(m^k), y)}_{\text{agreement term}} - \lambda \underbrace{\sum_{k\in[1,M]} \mathbb{D}_{KL}(\Phi_{M/k}(x) \| \phi_k(m^k))}_{\text{KL-based reconcilement regularization term}} \\ &+ \alpha \underbrace{\sum_{k\in[1,M]} \mathcal{L}_{mcr}(-\nabla_{\phi_{k-1}} \ell(\phi_{k-1}(m^{k-1}), y), -\nabla_{\phi_k} \ell(\phi_k(m^k), y))}_{\text{MCR term}} \\ &+ \underbrace{\sum_{k\in[1,M]} \mathcal{L}(\Phi_M(x), y)}_{\text{GRS term}}\end{aligned} \tag{10}$$

The pseudo-code of training ReconBoost is detailed in Alg.1. In lines 4 to 7, we calculate the dynamic modality-specific loss including the agreement term, KL-based reconcilement regularization term, and MCR term to update the k-th modality learner. After updating, in lines 10 to 13, we employ the GRS to perform global rectification. The process then continues with the update of the next modality learner.

4. Experiments

In this section, we provide the empirical evaluation across a wide range of multi-modal datasets to show the superior per-

Algorithm 1: ReconBoost Algorithm

Input: Observations $\mathcal{D}_{train}$, iterations of each stage T, lr in alternating-boosting stage γ, lr in global rectification stage η
Output: Well-trained model Φ_M

repeat
  ▷ Alternating-boosting Strategy
  In round s, pick up a modality-specific learner $\phi_k, k \in [1, M]$ to be updated in order;
  for $t = 0$ ***to*** $T-1$ **do**
    Sample $\forall\{x_i, y_i\} \in \mathcal{D}_{train}$;
    Calculate modality-specific loss $\ell_A\left(\phi_k^t\right) = \ell(\phi_k^t(m_i^k), y_i) - \lambda \mathbb{D}_{KL,s} + \alpha \ell_{mcr}$;
    Update $\vartheta_k^{t+1} = \vartheta_k^t - \gamma \cdot \nabla_{\vartheta_k^t} \ell_A\left(\phi_k^t\right)$;
  Add the model ϕ_k^T into the $\Phi_{M/k}$, denoted as $\Phi_{M,s}$;
  ▷ Global Rectification Scheme
  for $t = 0$ ***to*** $T-1$ **do**
    Sample $\forall\{x_i, y_i\} \in \mathcal{D}_{train}$;
    Calculate loss: $\ell_G(\Phi_{M,s}^t) = \ell(\sum_{k=1}^M \phi_k^t(m_i^k), y_i)$;
    Update all modality learners $\forall m \in [1, M]$ $\vartheta_m^{t+1} = \vartheta_m^t - \eta \cdot \nabla_{\vartheta_m^t} \ell_G(\Phi_{M,s}^t)$;
until *converge*;
return Φ_M.

formance of ReconBoost. ***Due to space limitations, please refer to App.C and D for an extended version.***

4.1. Experimental Setup

Dataset Descriptions. We conduct empirical experiments on several common multi-modal benchmarks. Specifically, **AVE** (Tian et al., 2018) dataset is designed for audio-visual event localization and includes 28 event classes. **CREMA-D** (Cao et al., 2014) is an audio-visual video dataset for speech emotion recognition including 6 emotion classes. **ModelNet40** (Wu et al., 2015) a large-scale 3-D model dataset, with the front and rear view to classify the object, following (Wu et al., 2022) and (Du et al., 2023). **MOSEI** (Zadeh et al., 2018), **MOSI** (Zadeh et al., 2016), and **CH-SIMS** (Yu et al., 2020) are multi-modal sentiment analysis datasets including three modalities, namely audio, image, and text. We defer the detailed introductions of these datasets to App.C.1

Competitors. To demonstrate the effectiveness of the proposed method, we compare it with some recent multi-modal learning methods that focus on alleviating modality competition. These competitors include **G-Blending** (Wang et al., 2020a), **OGM-GE** (Peng et al., 2022), **PMR** (Fan et al., 2023), **UME** (Du et al., 2023) and **UMT** (Du et al., 2023).

Table 1. Performance comparisons on AVE, CREMA-D, and ModelNet40 in terms of Acc(%). In the MN40 dataset, following UMT (Du et al., 2023), we use different views, so there are no prediction results of uni-audio modality, denoted as '-'.

Method	AVE	CREMA-D	MN40
AudioNet	59.37	56.67	-
VisualNet	30.46	50.14	80.51
Concat Fusion	62.68	59.50	83.18
G-Blending	62.75	63.81	84.56
OGM-GE	62.93	65.59	85.61
PMR	64.20	66.10	86.20
UME	66.92	68.41	85.37
UMT	67.71	70.97	90.07
Ours	**71.35**	**79.82**	**91.78**

We also include the **Concat**enation fusion method and **Uni**-modal methods. Detailed explanations of these competitors will be provided in App.C.2.

Implementation Details. All experiments are conducted on GeForce RTX 3090 and all models are implemented with Pytorch (Paszke et al., 2017). Specifically, for the AVE, CREMA-D and ModelNet40 datasets, we adopt ResNet-18 (He et al., 2016) as the backbone and modify the input channel according to the size of different modalities. For MOSEI, MOSI, and SIMS datasets, we conduct experiments with fully customized multimodal features extracted by the MMSA-FET toolkit. The uni-modal models are similar to (Williams et al., 2018). We adopt SGD (Robbins & Monro, 1951) as the optimizer. Specific data preprocessing, network design and optimization strategies are provided in App.C.3.

4.2. Overall Performance

The experimental results are presented in Tab.1 and Tab.2. Our proposed methods consistently outperform all competitors significantly across all datasets, underscoring the efficacy of our approach. In Tab.1, within a dataset featuring two modalities, all multi-modal learning methods exhibit improvements compared to the naive concatenation fusion method. This observation confirms the existence of modality competition in multi-modal joint learning, demonstrating the effective alleviation of modality competition by the compared methods. Specifically, given a well-trained AudioNet, our method achieves the most significant improvements when incorporating the visual modality, which justifies that our method can effectively make the most of cross-modal information.

In contrast to prior studies, we also evaluate the effectiveness of various modulation strategies on tri-modality datasets. Some earlier strategies (Peng et al., 2022; Fan et al., 2023) focused on mitigating competition between two modalities

Table 2. Performance comparisons on MOSEI, MOSI, and CH-SIMS datasets in terms of Acc(%).

Method	MOSEI	MOSI	CH-SIMS
AudioNet	52.29	54.81	58.20
VisualNet	50.35	57.87	63.02
TextNet	66.41	75.94	70.45
Concat Fusion	66.71	76.23	71.55
G-Blending	66.93	76.45	71.55
OGM-GE	66.67	76.01	71.10
PMR	66.41	76.12	70.90
UME	63.88	76.97	71.77
UMT	67.04	75.80	71.55
Ours	**68.61**	**77.96**	**73.88**

Table 3. Performance comparisons on the AVE and CREMA-D datasets in terms of mAP(%).

Method	Overall	Audio	Visual
Concat Fusion	36.43	34.71	20.08
OGM-GE	38.50	36.59	24.42
PMR	39.34	36.97	25.10
UME	40.02	37.12	30.45
UMT	42.58	35.65	32.41
Ours	**60.52**	**40.58**	**54.26**

and lacked generalization to multiple modalities. For these methods, we test combinations of different modalities and select better models for presentation. Our approach treats a multi-modal learning framework as a generalized ensemble model and demonstrates robust generalization across multiple modalities.

To further demonstrate ReconBoost's adaptability in broader contexts, we applied it to the retrieval task, a crucial area within computer vision. We assessed its performance using the Mean Average Precision (MAP) metric on the CREMA-D as shown in Tab.3. The detailed comparison results are provided in App.D.1

Applicable to Other Fusion Schemes. Our ReconBoost framework can be easily combined with several decision-level fusion methods, such as QMF (Zhang et al., 2023) and TMC (Han et al., 2021). Additionally, we benchmark against two straightforward baselines, Learnable Weighting (LW) and Naive Averaging (NA). To ensure fairness, we also included complex feature-based fusion, specifically MMTM (Joze et al., 2020), into our main competitors: OGM-GE, PMR, and UMT. As shwon in Tab.4, our method consistently outperforms others, highlighting the potential of more flexible fusion strategies to enhance performance. This reaffirms the effectiveness of ReconBoost. The detailed comparison

Table 4. Performance comparisons on the AVE and CREMA-D dataset in terms of Acc(%) with different fusion strategies. † indicates that MMTM is applied.

Method	AVE	CREMA-D
OGM_GE †	66.14	69.83
PMR †	67.72	70.14
UMT †	70.16	74.35
Ours + *NA*	71.35	79.82
Ours + *LW*	72.40	80.11
Ours + *TMC*	72.96	80.68
Ours + *QMF*	**73.20**	**81.11**

Table 5. Performance of the encoders trained by Uni-modal, Concatenation fusion, OGM-GE, UMT, and Ours in terms of Acc(%). We evaluate the encoders of all methods by training linear classifiers on them.

Method	CREMA-D		AVE	
	Visual	Audio	Visual	Audio
Uni-train	50.14	56.67	30.46	59.37
Concat Fusion	26.81	54.86	23.96	55.47
OGM-GE	29.17	55.42	25.52	56.51
PMR	29.21	55.60	26.30	57.20
UMT	45.69	58.47	31.25	60.70
Ours	**73.01**	**60.23**	**39.06**	**61.20**

results and analysis are provided in App.D.6.

4.3. Quantitative Analysis

Modality-specific Encoder Evaluation. We evaluate the encoders of Concatenation fusion, OGM-GE, PMR, UMT, and Ours by training linear classifiers on top of them. As shown in Tab.5, in most methods, the dominant modality (audio modality) encoder can achieve comparable performance compared with its uni-modal counterpart, however, the weak modality (visual modality) encoder is far behind. Uni-modal information remains underutilized, and uni-modal features suffer corruption during joint training. For UMT, employing a uni-modal distillation strategy aids in exploiting sufficient uni-modal features, enabling some encoders of UMT to slightly outperform their uni-modal counterparts. However, distilled knowledge will be slightly corrupted in the fusion due to modality competition.

Compared to them, the encoders trained by our proposed method achieve significant improvements. Benefiting from the alternating-learning paradigm, ReconBoost can avoid modality competition and ensure sufficient exploitations of uni-modal features. Furthermore, the innovative reconcilement regularization term effectively leverages comple-

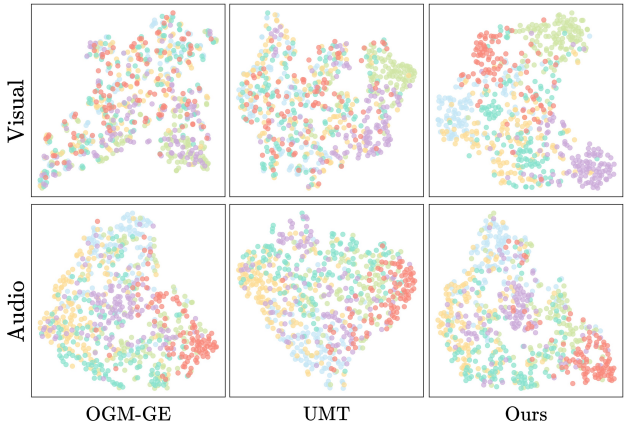


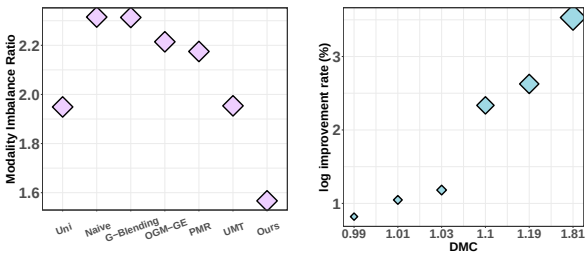
Figure 4. The visualization of the modality-specific feature among different competitors in the CREMA-D dataset by using the t-SNE method (Van der Maaten & Hinton, 2008).

mentary information between modalities. Our method's encoders achieve remarkable performance, which surpasses that of the uni-trained model.

Fig.4 shows the 2D embeddings of modality-specific features. In other methods, modality competition still exists. The features of the visual modality scatter randomly, reflecting low feature quality. Our approach focuses on improving the quality of latent features for each modality. Distinct clusters within each modality further highlight its effectiveness in reducing modality competition compared to other methods. The detailed comparison results are provided in App.D.9.

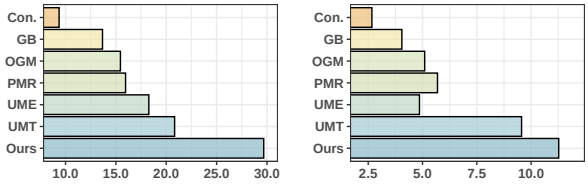
Modality Competition Analysis. Modality competition worsens the performance gap between modalities. To quantify this competition, we first measure the performance gap between modalities using the **modality imbalance ratio** (MIR). Moving further, we define the MIR of multi-modal learning methods relative to that of uni-modal learning as the degree of modality competition (DMC). Fig. 5(a) summarizes the modality imbalance ratio for all competitors on the AVE dataset. Although the MIR of various competitors is lower than that of naive concatenation fusion, it remains higher than the MIR under uni-modal learning, indicating the persistent challenge of modality competition. In contrast, our method effectively avoids modality competition, as revealed by the results. Fig.5(b) illustrates the DMC value of the concatenation fusion method across all datasets. Notably, as the degree of modality competition rises, so does the improvement our method offers. The in-depth analysis regarding the phenomena are shown in App.D.3.

Mutual Information Analysis. We quantify task-relevant mutual information (Liang et al., 2023b) in different multimodal models. Firstly, we decompose the mutual information into shared information and modality-specific unique



(a) MIR (b) DMC

Figure 5. Quantitative analysis of modality competition. (a) Modality imbalance ratio (MIR) for all competitors on the AVE dataset. (b) The correlation between the DMC in the concatenation fusion method and the improvement of our method is consistent across all datasets.



(a) Audio on CREMAD (b) Front View on MN40

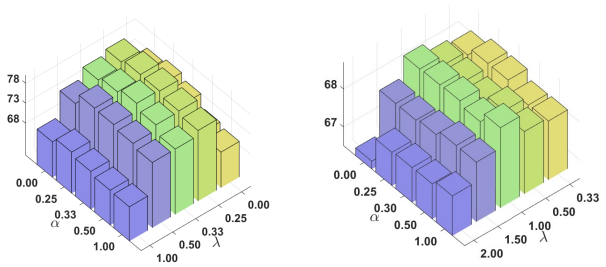
Figure 6. Quantitative analysis of task relevant mutual information in audio modality on the CREMA-D dataset and in front view on the MN40 dataset.

information. When only one modality $\mathcal{X}_1$ exists, the uni-modal only contains unique information. Then, in a multi-modal setting, maximize the information that $\mathcal{X}_2$ can bring becomes the key to improving the performance of multi-modal learning algorithms. We measure the information that $\mathcal{X}_2$ can bring using the difference in accuracy between using the multi-modal approach and the uni-modal model. As shown in Fig.6, we evaluate it among different competitors on two benchmarks and our method consistently outperforms others, demonstrating the potential of maximizing the useful information in each modality. The in-depth analysis are provided in App.D.4.

4.4. Ablation Study

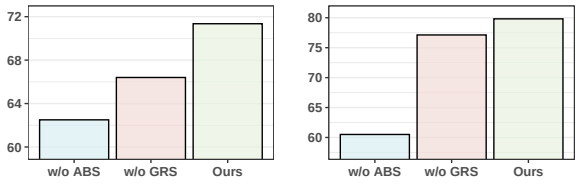
Sensitivity analysis of λ. Fig.7 demonstrates the performance of ReconBoost with varying λ on CREMA-D and MOSEI. We observe that a proper λ could extract complementary information and significantly improve the performance. Leveraging λ too aggressively may hurt the performance since excessive disagreement with others will damage modality-specific prediction accuracy. The detailed comparison results are provided in App.D.8.

Impact of Memory Consolidation Scheme. Fig.7 also explores the role of the α parameter in demonstrating the



(a) CREMA-D (b) MOSEI

Figure 7. Sensitivity analysis about λ and α on CREMA-D and MOSEI datasets.



(a) AVE Dataset (b) CREMA-D Dataset

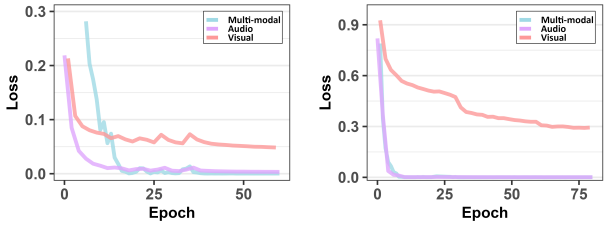
Figure 8. Abation study of global rectification scheme (GRS) on CREMA-D and AVE Dataset.

efficacy of the MCS. Keeping λ constant, we note that adjustments in α yield marginal yet meaningful improvements in performance. Given that λ primarily governs the level of agreement, its adjustment can significantly enhance memory consolidation in modality-specific learning. This suggests that while λ offers a broader range of manipulation for performance enhancement, fine-tuning with α allows for more precise and subtle improvements.

Effect of Global Rectification Scheme. Fig.8 illustrates the effectiveness of the global rectification scheme by comparing w/o GRS and Ours. GRS facilitates the optimization of the multi-modal learning objective, preventing the ensemble model from falling into unfavorable local minima. Even without GRS, our model achieves relatively good results, demonstrating that our alternating-boosting strategy effectively promotes the optimization of the objective.

4.5. Convergence

We present the convergence results on two benchmark datasets during the training process, including AVE and CREMA-D datasets. The performance results are shown in Fig.9. For ReconBoost, the updates of all modality learners are alternated, and the gradients across different modalities are naturally disentangled from each other. Therefore, the modality-specific loss curve descends without getting stuck.



(a) AVE Dataset (b) CREMA-D Dataset

Figure 9. Convergence results of ReconBoost on AVE and CREMA-D Dataset.

5. Conclusion

In this paper, we propose an effective multi-modal learning method based on an alternating learning paradigm to address the modality competition problem. Our method achieves a reconciliation between the exploitation of uni-modal features and the exploration of cross-modal interactions, with the crucial idea of incorporating a KL divergence based reconcilement regularization term. We have proven that optimizing modality-specific learners with this regularization is equivalent to the classic gradient-boosting algorithm. Therefore, the updated modality learner can fit the residual gap and promote the overall performance. We discard historical learners and only preserve the newest learners, forming an alternating-boosting strategy. Finally, the experiment results over a range of multi-modal benchmark datasets showcase significant performance improvements, affirming the effectiveness of the proposed method.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by the National Natural Science Foundation of China: 62236008, U21B2038, U23B2051, 61931008, 62122075, 61976202, 62206264 and 92370102, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680000, in part by the Innovation Funding of ICT, CAS under Grant No.E000000.

Impact Statement

We propose a general multi-modal learning method to deal with the bias toward weak modalities. When the weak modalities are sensitive to a potential group of people in society, it might be helpful to improve the overall fairness of the learning system.

References

Badirli, S., Liu, X., Xing, Z., Bhowmik, A., and Keerthi, S. S. Gradient boosting neural networks: Grownet. *ArXiv*, abs/2002.07971, 2020.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 59–66, 2018.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2019.

Cai, X., Nie, F., Huang, H., and Kamangar, F. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, pp. 1977–1984, 2011.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affective Comput.*, 5(4):377–390, 2014.

Chen, R. J., Lu, M. Y., Williamson, D. F., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40 (8):865–878, 2022.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *SIGKDD*, pp. 785–794, 2016.

Chen, X., Lin, K.-Y., Wang, J., Wu, W., Qian, C., Li, H., and Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, pp. 561–577, 2020.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. Covarep — a collaborative voice analysis repository for speech technologies. In *ICASSP*, pp. 960–964, 2014.

Deng, X. and Dragotti, P. L. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE TPAMI*, 43(10):3333–3348, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., and Zhao, H. On uni-modal feature learning in supervised multi-modal learning. In *ICML*, pp. 25, 2023.

Fan, Y., Xu, W., Wang, H., Wang, J., and Guo, S. Pmr: Prototypical modal rebalance for multimodal learning. In *CVPR*, pp. 20029–20038, 2023.

Feng, Y., Gao, Y., Zhao, X., Guo, Y., Bagewadi, N., Bui, N.-T., Dao, H., Gangisetty, S., Guan, R., Han, X., et al. Shrec'22 track: Open-set 3d object retrieval. *Computers & Graphics*, 107:231–240, 2022.

Freund, Y. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In *ICML*, pp. 148–156, 1996.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., and Scaramuzza, D. Event-based vision: A survey. *IEEE TPAMI*, 44(1):154–180, 2022.

Gao, Y., Li, S., Li, Y., Guo, Y., and Dai, Q. Superfast: 200× video frame interpolation via event camera. *IEEE TPAMI*, 45(6):7764–7780, 2023.

Guan, D., Cao, Y., Liang, J., Cao, Y., and Yang, M. Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *ArXiv*, abs/1802.09972, 2018.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multi-view classification. In *ICLR*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hessel, J. and Lee, L. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *EMNLP*, pp. 861–877, 2020.

Hu, W., Miyato, T., Tokui, S., Matsumoto, E., and Sugiyama, M. Learning discrete representations via information maximizing self-augmented training. In *ICML*, pp. 1558–1567, 2017.

Huang, F., Ash, J., Langford, J., and Schapire, R. Learning deep resnet blocks sequentially using boosting theory. *ArXiv*, abs/1706.04964, 2018.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multimodal network fail in deep learning? (Provably). In *ICML*, pp. 9226–9259, 2022.

Ivanov, S. and Prokhorenkova, L. Boost then convolve: Gradient boosting meets graph neural networks. In *ICLR*, 2021.

Jiang, Y., Xu, Q., Yang, Z., Cao, X., and Huang, Q. Dm2c: Deep mixed-modal clustering. In *NeurIPS*, pp. 5880–5890, 2019.

Jiang, Y., Hua, C., Feng, Y., and Gao, Y. Hierarchical set-to-set representation for 3-d cross-modal retrieval. *IEEE TNNLS*, pp. 1–13, 2023a.

Jiang, Y., Wang, Y., Li, S., Zhang, Y., Zhao, M., and Gao, Y. Event-based low-illumination image enhancement. *IEEE TMM*, pp. 1–12, 2023b.

Jing, L., Vahdani, E., Tan, J., and Tian, Y. Cross-modal center loss for 3d cross-modal retrieval. In *CVPR*, pp. 3142–3151, 2021.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. Mmtm: Multimodal transfer module for cnn fusion. In *CVPR*, pp. 13289–13299, 2020.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Li, J., Qiang, W., Zheng, C., Su, B., Razzak, F., Wen, J.-R., and Xiong, H. Modeling multiple views via implicitly preserving global consistency and local complementarity. *IEEE TKDE*, 2022.

Li, Z., Tang, J., and Mei, T. Deep collaborative embedding for social image understanding. *IEEE TPAMI*, 41(9): 2070–2083, 2019.

Liang, P. P., Deng, Z., Ma, M. Q., Zou, J., Morency, L.-P., and Salakhutdinov, R. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, 2023a.

Liang, P. P., Deng, Z., Ma, M. Q., Zou, J. Y., Morency, L.-P., and Salakhutdinov, R. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, pp. 32971–32998, 2023b.

Liang, P. P., Lyu, Y., Chhablani, G., Jain, N., Deng, Z., Wang, X., Morency, L.-P., and Salakhutdinov, R. Multiviz: Towards visualizing and understanding multimodal models. In *ICLR*, 2023c.

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

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. In *NeurIPS*, pp. 14200–14213, 2021.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

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

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Ruan, D., Ji, S., Yan, C., Zhu, J., Zhao, X., Yang, Y., Gao, Y., Zou, C., and Dai, Q. Exploring complex and heterogeneous correlations on hypergraph for the prediction of drug-target interactions. *Patterns*, 2(12), 2021.

Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., and Gross, H.-M. Efficient rgb-d semantic segmentation for indoor scene analysis. *ArXiv*, abs/2011.06961, 2021.

Shahroudy, A., Ng, T.-T., Gong, Y., and Wang, G. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE TPAMI*, 40(5):1045–1058, 2017.

Shalev-Shwartz, S. Selfieboost: A boosting algorithm for deep learning. *ArXiv*, abs/1411.3436, 2014.

Shao, Z., Li, F., Zhou, Y., Chen, H., Zhu, H., and Yao, R. Identity-invariant representation and transformer-style relation for micro-expression recognition. *Applied Intelligence*, 53(17):19860–19871, 2023.

Shao, Z., Zhou, Y., Li, F., Zhu, H., and Liu, B. Joint facial action unit recognition and self-supervised optical flow estimation. *Pattern Recognition Letters*, 181:70–76, 2024.

Sun, K., Zhu, Z., and Lin, Z. Adagcn: Adaboosting graph convolutional networks into deep models. *ArXiv*, abs/1908.05081, 2019.

Tang, J., Shu, X., Qi, G.-J., Li, Z., Wang, M., Yan, S., and Jain, R. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE TPAMI*, 39(8):1662–1674, 2017.

Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. Audio-visual event localization in unconstrained videos. In *ECCV*, pp. 247–263, 2018.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Wan, Z., Mao, Y., Zhang, J., and Dai, Y. Rpeflow: Multimodal fusion of rgb-pointcloud-event for joint optical flow and scene flow estimation. In *ICCV*, pp. 10030–10040, 2023.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *CVPR*, pp. 12692–12702, 2020a.

Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., and Huang, J. Deep multimodal fusion by channel exchanging. In *NeurIPS*, pp. 4835–4845, 2020b.

Wang, Y., Zhang, Y., Guo, Q., Zhao, M., and Jiang, Y. Rnve: A real nighttime vision enhancement benchmark and dual-stream fusion network. *IEEE Signal Process Lett.*, 31: 131–135, 2024.

Wei, Y., Hu, D., Tian, Y., and Li, X. Learning in audio-visual context: A review, analysis, and new perspective, 2022.

Williams, J., Kleinegesse, S., Comanescu, R., and Radu, O. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language*, 2018. doi: 10.18653/v1/W18-3302.

Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pp. 24043–24055, 2022.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pp. 1912–1920, 2015.

Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., and Zhu, W. Avqa: A dataset for audio-visual question answering on videos. In *ACM MM*, pp. 3480–3491, 2022.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., and Yang, K. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *ArXiv*, abs/1606.06259, 2016.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Annual Meeting of the Association for Computational Linguistics*, 2018.

Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. Causal intervention for weakly-supervised semantic segmentation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *NeurIPS*, volume 33, pp. 655–666, 2020.

Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J. T., and Peng, X. Provable dynamic fusion for low-quality multimodal data. In *ICML*, pp. 17, 2023.

Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., and Zhang, C. Multimodal fusion on low-quality data: A comprehensive survey, 2024.

Contents

A. Prior Arts

In this section, we briefly review the closely related studies along with our main topic.

A.1. Multi-modal Learning

Recent decades have witnessed the development of multi-modal learning research which covers various fields like cross-modal retrieval (Jiang et al., 2023a; Feng et al., 2022), video frame interpolation (Gao et al., 2023), image reconstruction (Wang et al., 2024; Jiang et al., 2023b), visual question answering (Yang et al., 2022), and clustering (Jiang et al., 2019; Hu et al., 2017). Intuitively, multi-modal models integrate information from multiple sensors to outperform their uni-modal counterparts. For example, event cameras as new vision sensors can compensate for the shortcomings of standard cameras in the face of abnormal light conditions or challenging high-speed scenarios (Gallego et al., 2022). These examples underscore the effectiveness of multi-modal approaches in addressing specific challenges and highlight the advantages arising from the fusion of diverse sensor modalities.

Numerous studies (Liang et al., 2023c; Du et al., 2023; Liang et al., 2023c; Hessel & Lee, 2020; Zhang et al., 2023; Li et al., 2022; Liang et al., 2023a) mainly concentrate on integrating modality-specific features into a shared representation for diverse tasks. Employed fusion methods encompass early/intermediate fusion (Seichter et al., 2021; Nagrani et al., 2021) as well as late fusion (Peng et al., 2022; Fan et al., 2023; Du et al., 2023; Wang et al., 2020a). Recent intermediate fusion methods utilize attention mechanisms that connect multi-modal signals during the modality-specific feature learning stage (Nagrani et al., 2021). While intermediate fusion may enhance representation learning, late fusion consistently stands out as the most prevalent and widely used approach, owing to its interpretability and practicality. Evolving from the naive late-fusion method, more methods are using dynamic fusion (Guan et al., 2018; Zhang et al., 2023) approaches to unleash the value of each modality and reduce the impact of low-quality multi-modal data.

A.2. Balanced Multi-modal Learning

However, recent theoretical evidence (Huang et al., 2022) illustrated that current paradigms of multi-modal learning encounter ***Modality Competition.*** Such a problem occurs when the objective for different modalities is optimized synchronously. In optimization, the modality with faster convergence dominates the learning process. Therefore, the learning parameters of other modalities can not be updated in a timely and effective manner. It will limit the optimization of the uni-modal branch and cannot fully exploit the information of the uni-modal, becoming a bottleneck in the performance of multi-modal learning.

To fill this gap, several studies (Wang et al., 2020a; Peng et al., 2022; Du et al., 2023; Fan et al., 2023) are proposed to balance the optimization process across different modality learners and promote the uni-modal learning. G-Blending (Wang et al., 2020a) incorporates uni-modal classifiers with extra supervised signals in multi-modal learning to effectively blend modalities. OGM-GE (Peng et al., 2022) focuses on suppressing the dominant modality and assisting the training of others through adaptive gradient modulations. PMR (Fan et al., 2023) employs the prototypical cross-entropy loss to accelerate the learning process of the weak modality. Additionally, UMT (Du et al., 2023) distills knowledge from well-trained uni-modal models in multi-modal learning, which can effectively benefit from uni-modal learning. In general, the majority of prior studies adopt a synchronous learning paradigm.

A.3. Boosting

Boosting is a commonly used learning approach in machine learning (Friedman, 2001; Freund, 1995; Friedman, 2001; Freund et al., 1996; Freund & Schapire, 1997). It enhances the performance of a basic learner by combining multiple weaker learners. In each iteration of boosting, the weaker learner focuses on the residual between the truth and its estimation. Decision trees are the most common weak learners that are used in boosting frameworks. Popular boosting algorithms include AdaBoost (Freund & Schapire, 1997), GBDT(Friedman, 2001), and XGBoost (Chen & Guestrin, 2016).

Inspired by the success of boosting in machine learning, boosting has recently received research attention in the deep learning community. Unlike traditional methods to construct ensembles of learners, SelfieBoost (Shalev-Shwartz, 2014) boosts the accuracy of a single network while discarding intermediate learners. (Huang et al., 2018) builds a ResNet-style architecture based on multi-channel telescoping sum boosting theory. AdaGCN (Sun et al., 2019) interprets a multi-scale graph convolutional network as an ensemble model and trains it using AdaBoost. BGNN (Ivanov & Prokhorenkova, 2021) combines the GBDT and GNN by iteratively adding new trees that fit the gradient updates of GNN.

B. Proof of Theorem 3.1

Restate of Theorem 3.1. When the reconcilement regularization satisfies,

$$\lambda \cdot \nabla_{\phi_k} \mathbb{D}_s \left(\Phi_{M/k}(x_i), \phi_k(\vartheta_k; m_i^k) \right) = \nabla_{\phi_k} \ell \left(\phi_k(\vartheta_k; m_i^k), y_i \right) - \nabla_{\phi_k} \ell \left(\phi_k(\vartheta_k; m_i^k), -\nabla_{\Phi_{M/k}} \ell(\Phi_{M/k}(x_i), y_i) \right)$$

It leads to equivalent optimization goals:

$$\nabla_{\vartheta_k} \tilde{\mathcal{L}}_s \left(\phi_k(m^k), y \right) \iff \nabla_{\vartheta_k} \mathcal{L} \left(\phi_k(m^k), -\nabla_{\Phi_{M/k}} \ell(\Phi_{M/k}(x), y) \right)$$

Proof. By using the right-hand side as the objective for gradient boosting, the k-th modality learner's parameters update as follows:

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \frac{1}{N} \sum_{i=1}^{N} \ell \left(\phi_k \left(\vartheta_k^t; m_i^k \right), -\nabla_{\Phi_{M/k}} \ell \left(\Phi_{M/k}(x_i), y_i \right) \right) \tag{11}$$

$$= \vartheta_k^t - \eta_1 \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial \phi_k(\vartheta_k^t; m_i^k)}{\partial \vartheta_k^t} \right)^T \nabla_{\phi_k} \ell \left(\phi_k \left(\vartheta_k; m_i^k \right), -\nabla_{\Phi_{M/k}} \ell \left(\Phi_{M/k}(x_i), y_i \right) \right) \tag{12}$$

If the left-hand side is the optimization strategy, then the objective becomes:

$$\tilde{\mathcal{L}}_s(\phi_k(m^k), y) = \frac{1}{N} \sum_{i=1}^{N} \left[\ell \left(\phi_k \left(\vartheta_k; m_i^k \right), y_i \right) - \lambda \cdot \underbrace{\mathbb{D}_s \left(\Phi_{M/k}(x_i), \phi_k \left(\vartheta_k; m_i^k \right) \right)}_{\text{reconcilement regularization term}} \right] \tag{13}$$

Through gradient optimization, we update the k-th modality learner's parameters:

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \tilde{\mathcal{L}}_s \left(\phi_k(m^k), y \right) \tag{14}$$

$$= \vartheta_k^t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial \phi_k \left(\vartheta_k^t; m_i^k \right)}{\partial \vartheta_k^t} \right)^T \cdot \left[\nabla_{\phi_k} \ell \left(\phi_k(\vartheta_k; m_i^k), y_i \right) - \lambda \cdot \nabla_{\phi_k} \mathbb{D}_s \left(\Phi_{M/k}(x_i), \phi_k \left(\vartheta_k; m_i^k \right) \right) \right] \tag{15}$$

$$= \vartheta_k^t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial \phi_k \left(\vartheta_k^t; m_i^k \right)}{\partial \vartheta_k^t} \right)^T \nabla_{\phi_k} \ell \left(\phi_k \left(\vartheta_k; m_i^k \right), -\nabla_{\Phi_{M/k}} \ell \left(\Phi_{M/k} \left(x_i \right), y_i \right) \right) \tag{16}$$

Thus, we conclude that

$$\nabla_{\vartheta_k} \tilde{\mathcal{L}}_s \left(\phi_k(m^k), y \right) \iff \nabla_{\vartheta_k} \mathcal{L} \left(\phi_k(m^k), -\nabla_{\Phi_{M/k}} \ell \left(\Phi_{M/k}(x), y \right) \right) \tag{17}$$

□

Here, to better understand the generality of our method and theorem, we will consider the case where the optimization loss function is Cross Entropy loss. CE loss is widely used for problems like classification, retrieval, and contrastive learning. As a corollary of the theorem we have:

Restate of Corollary 3.2. Let the reconcilement regularization be a KL divergence (Kullback & Leibler, 1951) function:

$$\mathbb{D}_s \left(\Phi_{M/k} \left(x_i \right), \phi_k \left(\vartheta_k; m_i^k \right) \right) = \mathbb{D}_{KL,s} \left(\Phi_{M/k} \left(x_i \right) \| \phi_k \left(\vartheta_k; m_i^k \right) \right)$$

Then,

$$\nabla_{\vartheta_k} \tilde{\mathcal{L}}_s \left(\phi_k \left(m^k \right), y \right) \iff \nabla_{\vartheta_k} \mathcal{L} \left(\phi_k(m^k), -\nabla_{\Phi_{M/k}} \ell \left(\Phi_{M/k}(x), y \right) \right)$$

where ℓ is the CE loss.

Proof. By using the right-hand side as the objective for gradient boosting, the k-th modality learner's parameters can be updated as:

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \frac{1}{N} \sum_{i=1}^{N} \ell(\phi_k(\vartheta_k^t; m_i^k), \tilde{y}_i) \tag{18}$$

The pseudo-label $\tilde{y}_i$ is

$$\tilde{y}_i = -\frac{\partial \ell(\Phi_{M/k}(x_i), y_i)}{\partial \Phi_{M/k}(x_i)} = y_i - \sigma_i \tag{19}$$

$\sigma_i \in \mathbb{R}^Y$ means the prediction score for i-th sample.

Therefore, we have

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{Y} (\sigma_{i,j} - c_{i,j}) \cdot \log(\rho_{i,j}^{k,t}) \tag{20}$$

Here, ρ_i^k is the prediction of the k-th modality learner on the i-th sample.

The left-hand objective is

$$\tilde{\mathcal{L}}_s(\phi_k(m^k), y) = \frac{1}{N} \sum_{i=1}^{N} \left[\ell\left(\phi_k\left(\vartheta_k; m_i^k\right), y_i\right) - \lambda \cdot \underbrace{\mathbb{D}_{KL,s}\left(\Phi_{M/k}\left(x_i\right) \| \phi_k\left(\vartheta_k; m_i^k\right)\right)}_{\text{KL-based reconcilement regularization}} \right] \tag{21}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[-\sum_{j=1}^{Y} c_{i,j} \log\left(\rho_{i,j}^k\right) - \lambda \cdot \sum_{j=1}^{Y} \sigma_{i,j} \ln \frac{\sigma_{i,j}}{\rho_{i,j}^k} \right] \tag{22}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[-\sum_{j=1}^{Y} c_{i,j} \log\left(\rho_{i,j}^k\right) + \lambda \cdot \sum_{j=1}^{Y} \sigma_{i,j} \ln \rho_{i,j}^k - \lambda \cdot \sum_{j=1}^{Y} \sigma_{i,j} \ln \sigma_{i,j} \right] \tag{23}$$

Through gradient optimization, in t-th iteration, the parameters of the k-th modality learner can be updated as:

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \tilde{\mathcal{L}}_s(\phi_k(m^k), y) \tag{24}$$

$$= \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{Y} \left(\lambda \cdot \sigma_{i,j} \ln \rho_{i,j}^k - c_{i,j} \log(\rho_{i,j}^k)\right) \tag{25}$$

Thus, with specific λ, we can reach the conclusion. □

C. Additional Experiment Setting

In this section, we elaborate on the setup of the main experiment, including dataset description, several state-of-the-art baselines, and implementation details.

C.1. Dataset Description

We perform empirical studies on six public benchmark datasets, including:

- **AVE**[1] (Tian et al., 2018). The AVE dataset is designed for audio-visual event localization. The dataset contains 4143 videos covering 28 event categories and videos in AVE are temporally labeled with audio-visual event boundaries. Each video contains at least one 2s long audio-visual event. The dataset covers a wide range of audiovisual events from different domains, such as, human activities, animal activities, music performances, and vehicle sounds. All videos are collected from YouTube. The training and testing split of the dataset follows (Tian et al., 2018).

[1] https://sites.google.com/view/audiovisualresearch

- **CREMA-D**[2] (Cao et al., 2014). The CREMA-D dataset is an audio-visual video dataset for speech emotion recognition, which consists of 7442 original clips of 2-3 seconds from 91 actors speaking several short words. It comprises six different emotions: anger, disgust, fear, happy, neutral, and sad. Categorical emotion labels were collected using crowd-sourcing from 2443 raters. The training and testing split of the dataset follows the split (Cao et al., 2014).
- **ModelNet40**[3] (Wu et al., 2015). The ModelNet40 is from a large-scale 3-D CAD model dataset ModelNet for object classification. ModelNet40 is a subset of ModelNet, which contains 40 popular object categories. We use the front view and the rear view to classify the 3-D object, following (Wu et al., 2022) and (Du et al., 2023). The dataset split for training and testing follows the standard protocol as described in (Wu et al., 2015).
- **MOSI**[4] (Zadeh et al., 2016). The CMU-MOSI dataset is one of the most popular benchmark datasets for multi-modal sentiment analysis (MSA). It comprises 2199 short monologue video clips taken from 93 Youtube movie review video. Human annotators label each sample with a sentiment score from -3 (strongly negative) to 3 (strongly positive). We view this as a three classification problem, with the categories being negative, neutral, and positive. The training and testing split of the dataset follows the split (Zadeh et al., 2016).
- **MOSEI**[4] (Zadeh et al., 2018). The CMU-MOSEI dataset expands its data with a higher number of utterances, greater variety in samples, speakers, and topics over CMU-MOSI.The dataset contains 23453 annotated video utterances, from 5000 videos, 1000 distinct speakers and 250 different topics. The training and testing split of the dataset follows the split (Zadeh et al., 2018).
- **CH-SIMS**[4] (Yu et al., 2020). The SIMS dataset is a Chinese MSA benchmark with fine-grained annotations of modality. The dataset consists of 2281 refined video segments collected from different movies, TV serials, and variety shows with spontaneous expressions, various head poses, occlusions, and illuminations. Human annotators label each sample with a sentiment score from -1 (strongly negative) to 1 (strongly positive). We treat this as a three classification problem, with the categories being negative, neutral, and positive. The training and testing split of the dataset follows the split (Yu et al., 2020).

To summarize, the overall statistical information is included in Tab.6.

Table 6. The statistics of all datasets used in the experiments.

Dataset	Task	# Train	# Test	# Category	Modality		
					Audio	Visual	Text
CREMAD	Speech emotion recognition	6698	744	6	✓	✓	✗
AVE	Event localization	3339	402	28	✓	✓	✗
ModelNet40	Object classification	9438	2468	40	✗	✓	✗
MOSEI	Emotion recognition	16327	4659	3	✓	✓	✓
MOSI	Emotion recognition	1284	686	3	✓	✓	✓
SIMS	Emotion recognition	1368	457	3	✓	✓	✓

C.2. Competitors

We compare the performance of our proposed method with several state-of-the-art baselines, including:

- **G-Blending** (Wang et al., 2020a) proposes Gradient Blending to obtain an optimal blending of modalities based on their over-fitting behaviors.
- **OGM-GE**[5] (Peng et al., 2022) proposes on-the-fly gradient modulation to adaptively control the optimization of each modality, via monitoring the discrepancy of their contribution towards the learning objective.

[2] https://github.com/CheyneyComputerScience/CREMA-D
[3] https://modelnet.cs.princeton.edu/
[4] https://drive.google.com/drive/folders/1A2S4pqCHryGmiqnNSPLv7rEg63WvjCSk?usp=sharing
[5] https://github.com/GeWu-Lab/OGM-GE_CVPR2022

- **PMR**[6] (Fan et al., 2023) proposes the prototypical modal rebalance strategy to address the modality imbalance problem, accelerating the slow modality with prototypical cross entropy loss and reducing the inhibition from dominant modality with prototypical entropy regularization term.
- **UME**[7] (Du et al., 2023) weights the predictions of well-trained uni-modal model directly.
- **UMT**[7] (Du et al., 2023) distills the well-trained uni-modal features to the corresponding parts of multi-modal late-fusion models and fusion the multi-modal features to obtain the final score.

C.3. Implementation Details

C.3.1. Network architecture

With respect to AVE, CREMA-D, and ModelNet40 datasets, the ResNet18 (He et al., 2016) is adopted as the backbone. **For AVE,** 3 frames of size $224 \times 224 \times 3$ are uniformly sampled from each 10-second clip as visual input and the whole audio data is transformed into a spectrogram of size 257×1004 by librosa[8] using a window with a length of 512 and overlap of 353. **For CREMA-D,** 1 frame of size $224 \times 224 \times 3$ is extracted from each video clip, and audio data is transformed into a spectrogram of size 257×299 with a length of 512 and overlap of 353. **In ModelNet40,** we resize the input front and rear views of a 3D object and CenterCrop it to $224 \times 224 \times 3$. To make the model compatible with the different data modalities mentioned above, we modify the input channel of ResNet-18 while keeping the remaining parts unchanged. Specifically, it takes the images as inputs and generates 512 dimension features, and takes the audio as inputs and outputs 512 dimension features, respectively. Then, a fully connected layer is established on top of the backbone model to make modality-specific predictions. Finally, multi-modal predictions are merged to obtain the final score.

For MOSEI, MOSI, and SIMS datasets, we conduct experiments with fully customized multimodal features extracted by the MMSA-FET[9] toolkit. The language features are extracted from pre-trained Bert(Devlin et al., 2018) and the pre-trained feature dimensions are 768 for all three datasets. Both MOSI and MOSEI use Facet[10] and SIMS uses the MultiComp OpenFace2.0 toolkit(Baltrusaitis et al., 2018) to extract facial expression features. The pre-trained visual feature dimensions are 20 for MOSI, 35 for MOSEI, and 709 for SIMS. Both MOSI and MOSEI extract acoustic features from COVAREP(Degottex et al., 2014) and SIMS uses LibROSA(McFee et al., 2015) speech toolkit with default parameters to extract acoustic features at 22050Hz. The pre-trained audio feature dimensions are 74 for MOSEI, 5 for MOSI, and 33 for SIMS. For these three datasets, we feed the pre-trained features into modality-specific backbones to extract the latent feature, with the hidden dimension set to 128. Following (Williams et al., 2018), the AudioNet and VisualNet are composed of three fully connected layers and the TextNet uses LSTM to capture long-distance dependencies in a text sequence. Then, a fully connected layer is established on top of the backbone model to make modality-specific predictions. Finally, multi-modal predictions are merged to obtain the final score.

C.3.2. Training details

All experiments are conducted on a Ubuntu 20.04 LTS server equipped with Intel(R) Xeon(R) Gold 5218 CPU@2.30GHz and RTX 3090 GPUs, and we implement all algorithms with PyTorch (Paszke et al., 2017). We adopt SGD (Robbins & Monro, 1951) as the optimizer and set the same learning rate in the alternating-boosting stage and rectification stage. The learning rate is 0.01 initially and multiplies 0.1 every 30 stages for the CREMA-D dataset, while multiplies 0.5 after 40 stages for the AVE dataset. For MOSEI, MOSI, CH-SIMS, and ModelNet40, the learning rate is 0.01 and remains constant. In one alternating-boosting stage, we will pick one modality learner to update and this modality learner will experience $T1$ epochs. Then, we will step into global rectification stage and the model will experience $T2$ epochs. $T1$ and $T2$ will vary depending on the datasets. In AVE, CREMA-D, ModelNet40, MOSEI, MOSI and SIMS, $T1$ is 4, 4, 4, 1, 1, 1 and $T2$ is 4, 4, 4, 1, 1, 1 respectively.

D. Additional Experiment Analysis

In this section, we provide additional experimental results and analysis to further support the conclusions in the main text.

[6] https://github.com/fanyunfeng-bit/Modal-Imbalance-PMR
[7] https://openreview.net/forum?id=mb7VM83DkyC
[8] https://librosa.org/
[9] https://github.com/thuiar/MMSA-FET
[10] https://imotions.com/products/imotions-lab/

Table 7. Performance comparisons on the AVE and CREMA-D datasets in terms of mAP(%).

Method	AVE			CREMA-D		
	MAP	Audio MAP	Visual MAP	MAP	Audio MAP	Visual MAP
Concat Fusion	35.25	37.23	18.82	36.43	34.71	20.08
OGM-GE	36.92	35.43	20.04	38.50	36.59	24.42
PMR	36.75	35.71	20.32	39.34	36.97	25.10
UME	34.91	33.41	21.93	40.02	37.12	30.45
UMT	36.72	34.64	21.76	42.58	35.65	32.41
CMCL	40.21	38.15	23.41	53.31	38.31	41.25
HSR	41.49	39.21	24.01	55.22	39.20	44.67
Ours	**43.85**	**42.71**	**25.22**	**60.52**	**40.58**	**54.26**

Table 8. Performance comparisons on the CREMA-D dataset in terms of Acc(%) when 50% of the image data is corrupted with Gaussian noise i.e., zero mean with the variance of σ^2.

Method	$\sigma^2 = 0.0$	$\sigma^2 = 0.1$	$\sigma^2 = 0.3$	$\sigma^2 = 0.5$	$\sigma^2 = 1.0$
VisualNet	50.14	47.42	43.71	40.30	35.35
Concat Fusion	59.50	58.70	58.13	57.70	57.10
OGM-GE	65.59	64.20	62.50	61.70	60.17
PMR	66.10	65.58	63.39	62.10	61.08
UME	68.41	66.49	63.28	62.04	61.46
UMT	70.97	68.76	64.92	63.01	62.23
Ours	**79.82**	**74.75**	**68.26**	**65.73**	**63.95**

D.1. Performance on Retrieval Task

To further demonstrate ReconBoost's adaptability in broader contexts, we apply it to the retrieval task, a crucial area within computer vision. For this purpose, we employ modality-specific pre-trained encoders to obtain latent features from each modality. For modality-specific retrieval, we utilize the respective latent features, whereas, for holistic retrieval, we combine all latent features to make predictions. Cosine similarity served as the metric for our retrieval scores. We assess its performance using the Mean Average Precision (MAP) metric on the CREMA-D and AVE datasets, as detailed in the subsequent Tab.7.

Additionally, we benchmark our approach against recent advancements in retrieval tasks. CMCL(Jing et al., 2021) introduces a cross-modal center loss for learning distinctive and modality-invariant features, showing impressive results in both in-domain and cross-modal retrieval. HSR(Jiang et al., 2023a) develops a hierarchical representation strategy, utilizing hierarchical similarity for retrieval tasks.

These comparisons reveal that the challenge of modality competition persists in retrieval tasks. However, ReconBoost effectively mitigates this issue, leading to superior performance.

D.2. Robustness Performance

Our initial learning approach assumes all modalities are of high quality. To assess how our method handles noisy data, we introduce Gaussian noise into different modalities and evaluate the performance using the CREMA-D dataset.

Case 1: In scenarios where 50% of the image data is distorted with Gaussian noise $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ $(\mu = 0)$, we observe the outcomes across various levels of noise intensity σ, as detailed in the Tab.8.

Case 2: Similarly, when 50% of the audio data encounters the same type of noise distortion, we document the performance

Table 9. Performance comparisons on the CREMA-D dataset in terms of Acc(%) when 50% of the audio data is corrupted with Gaussian noise i.e., zero mean with the variance of σ^2.

Method	$\sigma^2 = 0.0$	$\sigma^2 = 0.1$	$\sigma^2 = 0.3$	$\sigma^2 = 0.5$	$\sigma^2 = 1.0$
AudioNet	56.67	51.70	49.20	46.70	44.80
Concat Fusion	59.50	57.01	55.56	54.74	52.27
OGM-GE	65.59	63.28	62.09	59.56	56.49
PMR	66.10	63.74	62.83	60.29	57.10
UME	68.41	63.01	61.64	60.83	58.57
UMT	70.97	66.29	64.71	63.40	60.37
Ours	79.82	73.65	68.19	65.05	63.24

Table 10. Performance comparisons on the CREMA-D dataset in terms of Acc(%) when 50% of the image data and the audio data are corrupted with Gaussian noise i.e., zero mean with the variance of σ^2.

Method	$\sigma^2 = 0.0$	$\sigma^2 = 0.1$	$\sigma^2 = 0.3$	$\sigma^2 = 0.5$	$\sigma^2 = 1.0$
AudioNet	56.67	51.70	49.20	46.70	44.80
VisualNet	50.14	47.42	43.71	40.30	35.35
Concat Fusion	59.50	55.31	52.34	49.42	47.23
OGM-GE	65.59	60.14	57.36	54.26	50.38
PMR	66.10	62.57	58.19	55.14	51.25
UME	68.41	62.95	58.84	55.72	51.91
UMT	70.97	65.02	63.50	60.63	55.74
Ours	**79.82**	**71.83**	**67.17**	**63.09**	**57.60**

changes with different noise intensities σ, as shown in Tab.9.

Case 3: In cases where both audio and image data are 50% corrupted by Gaussian noise $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ $(\mu = 0)$, the impacts on performance with varying noise levels σ are summarized in the Tab.10.

Our observations indicate that despite the presence of noise, our method consistently outperforms competing approaches in all the scenarios mentioned above.

D.3. Modality Competition Analysis

In this subsection, we quantify the modality competition and analyze this phenomenon in more detail.

Given input data $\mathcal{X}$ that consists of M modalities, $\mathcal{X} = \{\mathcal{X}_1, \cdots, \mathcal{X}_M\}$, $\mathcal{Y}$ represents the ground-truth labels. We can train M separate encoders $\{\varphi_1^{uni}, \cdots, \varphi_M^{uni}\}$ and classifiers $\{f_1^{uni}, \cdots, f_M^{uni}\}$ for each modality through uni-modal training. We can also train encoders $\{\varphi_1^{mul}, \cdots, \varphi_M^{mul}\}$ for all modalities through multi-modal learning. Then, we build a classifier on the frozen modality-specific encoder, denoted as $\{f_1^{mul}, \cdots, f_M^{mul}\}$ for all modalities. $\texttt{Acc}(\cdot)$ represents the accuracy evaluation function.

For any two modalities $\mathcal{X}_i$ (the strong modality) and $\mathcal{X}_j$ (the weak modality), we define the **modality imbalance ratio (MIR)** in a uni-modal setting as:

$$\text{MIR}^{uni}(\mathcal{X}_i, \mathcal{X}_j) = \frac{\texttt{Acc}(f_i^{uni} \circ \varphi_i^{uni}(\mathcal{X}_i))}{\texttt{Acc}(f_j^{uni} \circ \varphi_j^{uni}(\mathcal{X}_j))} \tag{26}$$

In a multi-modal setting, the definition of MIR is:

$$\text{MIR}^{mul}(\mathcal{X}_i, \mathcal{X}_j) = \frac{\texttt{Acc}(f_i^{mul} \circ \varphi_i^{mul}(\mathcal{X}_i))}{\texttt{Acc}(f_j^{mul} \circ \varphi_j^{mul}(\mathcal{X}_j))} \tag{27}$$

MIR effectively measures the accuracy ratio between any two modalities, where a higher MIR indicates a more pronounced imbalance in learning across different modalities.

Furthermore, to assess the competition between multi-modal and uni-modal learning, we introduce the **Degree of Modality Competition (DMC)**. Specifically, DMC compares the MIR of a multi-modal learner to that of a uni-modal learner:

$$\text{DMC}(\mathcal{X}_i, \mathcal{X}_j) = \frac{\text{MIR}^{mul}(\mathcal{X}_i, \mathcal{X}_j)}{\text{MIR}^{uni}(\mathcal{X}_i, \mathcal{X}_j)} \tag{28}$$

A higher DMC value indicates more intense modality competition. We also expand DMC to accommodate three modalities by calculating the geometric mean of all modality pairs:

$$\text{DMC}(\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_k) = \sqrt[3]{\prod_{\substack{m,n\in\{i,j,k\}\\ m\neq n}} \text{DMC}(\mathcal{X}_m, \mathcal{X}_n)} \tag{29}$$

Tab.11 summarizes the modality imbalance ratio of different multi-modal learning methods on both the CREMA-D and AVE datasets. Tab.12 shows the DMC value of the concatenation fusion method across all datasets. Notably, as the degree of modality competition rises, so does the improvement our method offers.

Table 11. Modality imbalance ratio (MIR) and the degree of modality competition (DMC) for all competitors on the CREMA-D and AVE dataset. Audio modality is a strong modality.

Method	**CREMAD Dataset**				**AVE Dataset**			
	Audio	Visual	MIR	DMC	Audio	Visual	MIR	DMC
Uni-train	56.67	50.14	1.13	-	59.37	30.46	1.95	-
Concat Fusion	54.86	26.81	2.05	1.81	55.47	23.96	2.32	1.19
G-Blending	54.90	28.05	1.96	1.73	55.80	24.12	2.31	1.19
OGM-GE	55.42	29.17	1.90	1.68	56.51	25.52	2.21	1.14
PMR	55.60	29.21	1.90	1.68	57.20	26.30	2.17	1.12
UMT	58.47	45.69	1.28	1.13	60.70	31.07	1.95	1.00
Ours	**60.23**	**73.01**	**0.82**	**0.73**	**61.20**	**39.06**	**1.57**	**0.80**

Table 12. The correlation between the Degree of Modality Competition (DMC) using the concatenation fusion method and the enhancement of our method compared to that across all datasets. If the dataset lacks this modality, it is denoted as '-'.

Dataset	Uni-modal			Concat-fusion			DMC	Concat	Ours	Relative Improvement
	Audio	Visual	Text	Audio	Visual	Text				
CREMA-D	56.67	50.14	-	54.86	26.81	-	**1.81**	59.50	79.82	**34.15%**
AVE	59.37	30.46	-	55.47	23.96	-	**1.19**	62.68	71.35	**13.83%**
MOSEI	52.29	50.35	66.41	49.02	49.02	66.13	**1.01**	66.71	68.61	**2.85%**
MOSI	54.81	57.87	75.94	54.25	54.37	74.05	**0.99**	76.23	77.96	**2.27%**
CH-SIMS	58.20	63.02	70.45	54.27	59.74	68.71	**1.03**	71.55	73.88	**3.26%**

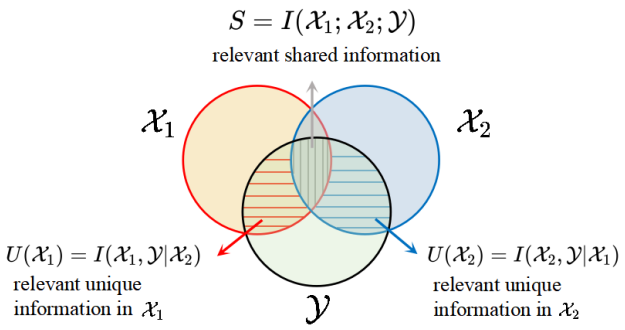


Figure 10. The visualization of the multi-modal information. This figure is derived from (Liang et al., 2023b).

D.4. Analysis of Mutual Information in Modalities

In this subsection, we will illustrate the effectiveness of our method from the perspective of mutual information. Assume that two modalities are denoted as $\mathcal{X}_1$ and $\mathcal{X}_2$. $\mathcal{Y}$ represents the groud-truth labels. Following (Liang et al., 2023b), we decompose the multi-modal information $I(\mathcal{X}_1, \mathcal{X}_2; \mathcal{Y})$ into three conditional mutual information (MI) terms and visualize the multi-modal information as Fig.10.

$$I(\mathcal{X}_1, \mathcal{X}_2; \mathcal{Y}) = \underbrace{I(\mathcal{X}_1; \mathcal{X}_2; \mathcal{Y})}_{S(\mathcal{X}_1,\mathcal{X}_2)=\text{relevant shared info.}} + \underbrace{I(\mathcal{X}_1, \mathcal{Y}|\mathcal{X}_2)}_{U(\mathcal{X}_1)=\text{relevant unique info. in } \mathcal{X}_1} + \underbrace{I(\mathcal{X}_2, \mathcal{Y}|\mathcal{X}_1)}_{U(\mathcal{X}_2)=\text{relevant unique info. in } \mathcal{X}_2} \tag{30}$$

When only one modality $\mathcal{X}_1$ exists, the uni-modal only contains unique information $U(\mathcal{X}_1)$. Then, in a multi-modal setting, maximize the information that $\mathcal{X}_2$ can bring becomes the key to improving the performance of multi-modal learning algorithms. We measure the information that $\mathcal{X}_2$ can bring using the difference in accuracy between using the multi-modal approach and the uni-modal model, as:

$$U(\mathcal{X}_2) + S(\mathcal{X}_1, \mathcal{X}_2) = \texttt{Acc}(\mathcal{X}_1, \mathcal{X}_2) - \texttt{Acc}(\mathcal{X}_1) \tag{31}$$

where $\texttt{Acc}(\mathcal{X}_1)$ and $\texttt{Acc}(\mathcal{X}_2)$ denote the accuracy of the unimodal learning algorithm using only $\mathcal{X}_1$ and $\mathcal{X}_2$ modalities, respectively; $\texttt{Acc}(\mathcal{X}_1, \mathcal{X}_2)$ denote the accuracy of the multi-modal learning algorithm using both $\mathcal{X}_1$ and $\mathcal{X}_2$ modalities.

Then, we evaluate it among different competitors on three benchmarks. Overall, as shown in Fig.11, our method consistently outperforms others, demonstrating the potential of maximizing the valuable information in each modality. This further illustrates the effectiveness of our method.

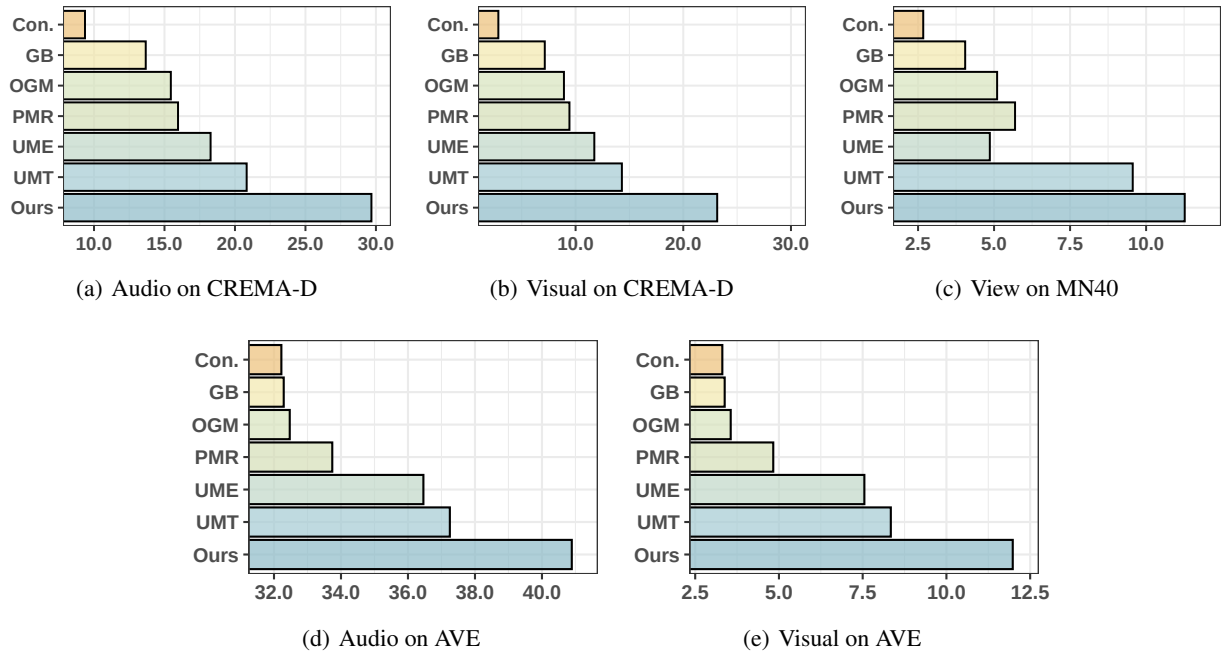
D.5. Modality Selection Strategy

In this subsection, we investigate the effect of different modality selection strategies. Our method expects to tackle the issue of modality competition. To this end, we alternate learning for each modality. This approach intuitively eases modality competition in gradient space by requiring separate use of modality-specific gradients. We now explore quality-guided criteria for modality selection. We introduce two additional modality selection schemes based on loss value as a measure of modality-specific data quality:

- S_1: We select the modality learner with the lowest loss value for updates in each round, favoring high-quality modalities.
- S_2: We select the modality learner with the highest loss value for updates in each round, prioritizing low-quality modalities.

We evaluate the effectiveness of these two optimization orders using the AVE dataset, as shown in Tab.13.

The results demonstrate that our method surpasses those based on quality selection. This might be because the S_1 strategy could cause low-quality modality learners to get stuck at poor local optima. Conversely, the S_2 strategy may restrict the potential of high-quality modalities.



(a) Audio on CREMA-D (b) Visual on CREMA-D (c) View on MN40

(d) Audio on AVE (e) Visual on AVE

Figure 11. Quantitative analysis of task relevant mutual information in modalities on the AVE, CREMA-D, and MN40 datasets.

Table 13. Performance comparisons on the AVE dataset in terms of Acc(%) with different selection strategies.

Optimization Order	Overall Acc	Audio Acc	Visual Acc
S_1	61.45	60.03	24.11
S_2	68.75	57.20	38.90
Ours	**71.35**	**61.20**	**39.06**

D.6. Applicable to Other Fusion Schemes

In this subsection, we investigate how to combine our approach with some multi-modal fusion methods to better improve performance. Multi-modal fusion methods are typically divided into two categories: feature-level and decision-level fusions (Baltrušaitis et al., 2019). Feature-level fusion combines latent features before making predictions, commonly used in multi-modal joint-learning approaches. In contrast, decision-level fusion aggregates predictions from each modality to reach a final decision. Our main paper demonstrates that joint learning can cause modality competition. To mitigate this, we introduced a new multi-modal learning framework based on decision-level fusion strategies, enhancing adaptability to complex decision-making scenarios. We modify our original decision aggregation formula as follows:

$$\Phi_M(x_i) = \sum_{k=1}^{M} w_k \cdot \phi_k(\vartheta_k; m_i^k) \tag{32}$$

where w_k signifies the importance of the k-th modality during inference.

Our ReconBoost framework can be easily combined with other fusion methods, particularly:

- **QMF** (Zhang et al., 2023) employs a dynamic, uncertainty-aware weighting mechanism at the decision level.
- **TMC** (Han et al., 2021) uses a dynamic approach to integrate modalities through the Dempster-Shafer theory efficiently.

Additionally, we benchmark against two straightforward baselines:

- Learnable Weighting (**LW**): Assigns trainable weights w_k to each modality and learns these weights alongside other parameters.
- Naive Averaging (**NA**): Averages predictions across modalities, setting $w_k = 1$ for all modalities.

Furthermore, to emphasize the superiority of our approach, we also evaluate a novel feature-based fusion competitor, MMTM (Joze et al., 2020), on the AVE and CREMA-D datasets.

Table 14. Performance comparisons on the AVE and CREMA-D dataset in terms of Acc(%) with different fusion strategies.

Method	AVE			CREMA-D		
	Overall Acc	Audio Acc	Visual Acc	Overall Acc	Audio Acc	Visual Acc
OGM_GE + MMTM	66.14	58.23	28.09	69.83	58.76	53.35
PMR + MMTM	67.72	58.47	28.73	70.14	58.94	54.23
UMT + MMTM	70.16	60.40	35.83	74.35	60.86	62.83
Ours + NA	71.35	61.20	39.06	79.82	60.23	73.01
Ours + LW	72.40	61.31	39.13	80.11	60.09	73.30
Ours + TMC	72.96	61.51	40.20	80.68	60.37	73.86
Ours + QMF	**73.20**	**61.96**	**40.85**	**81.11**	**60.94**	**73.87**

Overall, as shown in Tab.14, our method consistently outperforms others, highlighting the potential of more flexible fusion strategies to enhance performance. This reaffirms the effectiveness of ReconBoost.

D.7. Impact of Different Classifiers

In Tab.5, we limit classifiers to linear models to assess the effectiveness of our proposed ensemble method. Herein, we expand our evaluation to include non-linear classifiers featuring multiple fully connected (FC) layers and ReLU functions. Specifically, we develop non-linear classifier architecture, Fc+Relu+Fc, and FC+Relu+FC+Relu+FC, for the encoders used in all methods and test its performance on the CREMA-D dataset. As shown in Tab.15, we arrive at the conclusion that $1)$ modality competition exists no matter which classifier is used. $2)$ Our approach, ReconBoost, enhances the performance of encoders with various classifiers, demonstrating that our model effectively learns high-quality latent features.

D.8. Sensitivity Analysis of λ

To assess the impact of λ, we carry out additional sensitivity tests by altering λ's value. We present the results for CREMA-D, AVE, and ModelNet40 in the Tab.16. The performance of our method stays consistent with λ values between $[1/4, 1/2]$. Additionally, our method continues to achieve state-of-the-art results within this λ range.

Table 15. Performance comparisons on the CREMA-D dataset in terms of Acc(%) with different classifiers.

Method	FC		FC+Relu+FC		FC+Relu+FC+Relu+FC	
	Visual	Audio	Visual	Audio	Visual	Audio
Uni-train	50.14	56.67	50.25	56.97	50.31	57.10
Concat Fusion	26.81	54.86	26.89	54.91	26.96	55.02
OGM-GE	29.17	55.42	29.72	56.03	29.91	56.11
PMR	29.21	55.60	29.81	56.22	30.05	56.45
UMT	45.69	58.47	46.73	58.71	47.01	58.93
Ours	**73.01**	**60.23**	**73.34**	**60.24**	**73.86**	**60.37**

Table 16. Sensitivity analysis of λ on the CREMA-D, AVE and ModelNet40 datasets.

Method	AVE	CREMA-D	MN40
Concat Fusion	62.68	59.50	83.18
G-Blending	62.75	63.81	84.56
OGM-GE	62.93	65.59	85.61
PMR	64.20	66.10	86.20
UME	66.92	68.41	85.37
UMT	67.71	70.97	90.07
Ours $\lambda = 1/4$	**71.35**	79.26	91.13
Ours $\lambda = 1/3$	70.31	**79.82**	**91.78**
Ours $\lambda = 1/2$	69.53	77.13	91.25
Ours $\lambda = 1$	69.49	70.31	89.91

D.9. Latent Embedding Visualization

Taking a step further, we visualize the latent embedding of modality-specific features among different competitors on the CREMA-D datasets. Specifically, we first regard the outputs of the backbone as the latent vectors of images and then project them into a 2D case by t-SNE (Van der Maaten & Hinton, 2008). Comparing these results, we can see that our proposed method outperforms other competitors in all modalities since the cluster results of ReconBoost are more significant, especially in the weak modality. This again ascertains the advantages of our proposed approach.

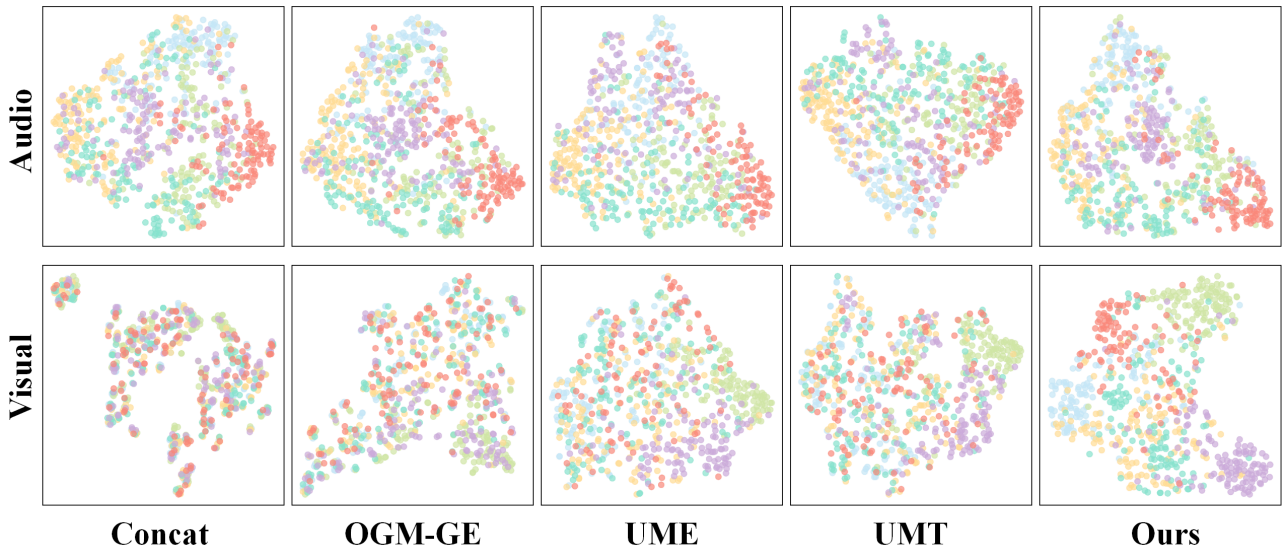


Figure 12. The visualization of the modality-specific feature among competitors in the CREMA-D dataset by using the t-SNE method (Van der Maaten & Hinton, 2008).