Deep Information-Balanced Multimodal Learning

Yang Qin, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu

Abstract—Multimodal learning aims to integrate diverse data sources to capture more comprehensive information about things, thus enhancing perception and understanding of the real world. However, inherent discrepancies between different modalities often lead to imbalanced optimization during multimodal learning, hindering performance improvement. To address this issue, in this paper, we present a Multimodal Information Balance (MIB) theory, grounded in Information Theory, to reveal that this imbalance arises from the imbalanced retention of complementary information during modality fusion, providing an intuitive and explainable perspective on the issue. Building on this insight, we propose a theoretical MIB criterion to adaptively balance the preservation of complementary information across individual modalities, thereby facilitating multimodal fusion. Using this criterion, we develop an Information-Balanced Multimodal Learning (IBML) framework to mine comprehensive and balanced multimodal information, achieving optimal learning. More specifically, IBML introduces Balance Information Optimization (BIO) module to maximize tractable lower bound objectives derived from the MIB criterion according to the optimization discrepancies across modalities, ensuring balanced retention of complementary information and enhancing information contributions during multimodal fusion. In addition, we present a supplementary and provable Task Complexity Modulation (TCM) module based on the MIB criterion to adjust task complexity discrepancies across input modalities, thus indirectly promoting the balanced preservation of complementary information throughout the learning process. Extensive experiments are conducted on eight multimodal datasets, spanning audio-visual recognition, image-text classification, and 2D-3D recognition, to verify the superiority and effectiveness of IBML. The code will be released publicly after in-peer review.

Index Terms—Multimodal learning, Multimodal balanced optimization, Multimodal fusion.

1 Introduction

Multimodal learning [1], [2], [3] has raised significant attention within the artificial intelligence community due to its potential to integrate complementary information across various modalities. As a substantial learning paradigm for human-like artificial intelligence, multimodal learning empowers machines to see, hear, and interact with the real world by jointly leveraging multiple sensory modalities, and has demonstrated excellent capabilities and promising progress in various domains [2], [4], [5], [6], [7]. However, due to inherent cross-modal and cross-model discrepancies, balancing the optimization of different modalities to extract complementary information from multiple modalities remains a significant challenge in realistic scenarios. To conquer this challenge, especially some recent studies [2], [8], [9] have revealed the issue of Imbalanced Multimodal Learning (IML) in multimodal joint learning, where certain modalities optimize faster than others, leading to a suboptimal performance. A widely accepted view for this stalemate is that deep models tend to prioritize learning from simpler (*strong*) modalities, causing them to dominate the optimization process while underutilizing more complex (*lazy*) modalities [2], [10], [11].

To tackle IML, various approaches have been proposed to coordinate the joint learning of different modalities [2], [11], [13], [14]. Existing methods [2], [8], [12] primarily focus on diminishing the dominance of strong modalities to balance optimization by heuristically adjusting the optimization pace of each modality. In contrast, some recent works [12], [15] aim to resolve modal optimization conflicts by preserving the inherent learning direction of lazy modalities while reducing the optimization intensity of strong ones, thus alleviating the IML issues. Despite promising

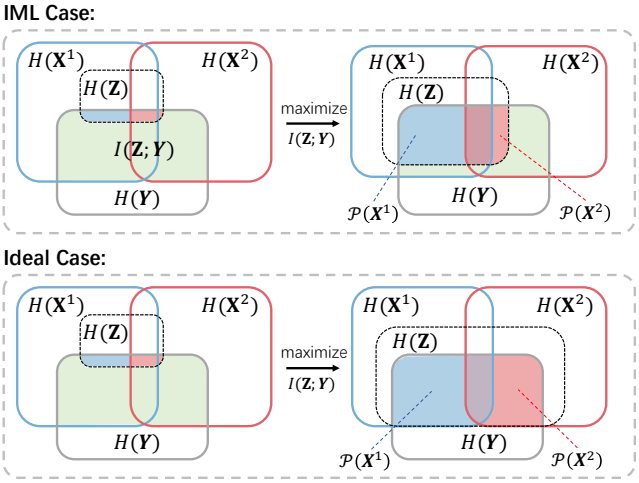
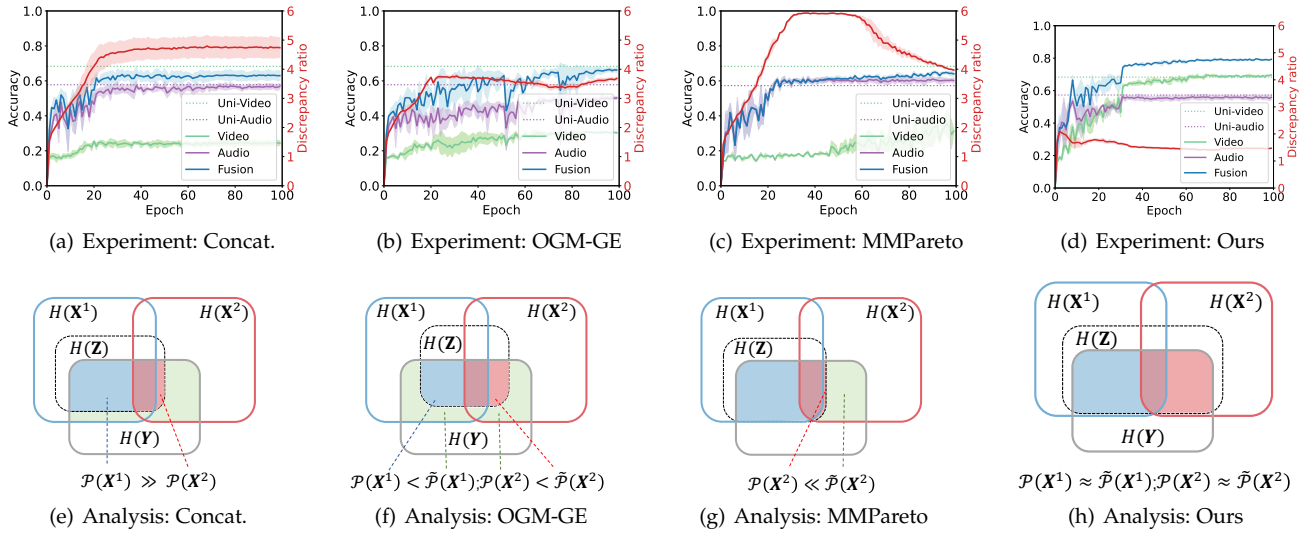


Fig. 1: The schematic illustration of maximizing the mutual information between two modalities and label semantics in Imbalanced Multimodal Learning (IML). $\mathbf{Z}$ is the fusion variable for modalities $\mathbf{X}^1$ and $\mathbf{X}^2$, $\mathbf{Y}$ is the labels, $H(\mathbf{X}^1)$ and $H(\mathbf{X}^2)$ are the entropy of the modal valuables. We reveal IML as the imbalance between complementary information ($\mathcal{P}(\mathbf{X}^1) = I(\mathbf{X}^1;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1,\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$ and $\mathcal{P}(\mathbf{X}^2) = I(\mathbf{X}^2;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1,\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$). We expect to maximize the complementary information impartially while maximizing the shared information $I(\mathbf{X}^1,\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$ (the ideal case). However, this is not well solved in existing methods.

Y. Qin, Y. Feng, Y. Sun, D. Peng, X. Peng, and P. Hu are with the College of Computer Science, Sichuan University, Chengdu, China, 610044.
1. Y. Qin and Y. Feng contributed equally.
2. Corresponding author: Peng Hu (e-mail:penghu.ml@gmail.com).



(a) Experiment: Concat. (b) Experiment: OGM-GE (c) Experiment: MMPareto (d) Experiment: Ours

(e) Analysis: Concat. (f) Analysis: OGM-GE (g) Analysis: MMPareto (h) Analysis: Ours

Fig. 2: Methods analysis. (a) - (d) show the modal discrepancy ratio and recognition accuracy of the vanilla concatenation fusion method (*i.e.*, Concat.), OGM-GE [2], MMPareto [12], and our method on the CREMA-D dataset [7]. The red solid line represents the discrepancy across modalities. The green, purple, and blue solid lines represent the video (*i.e.*, lazy modality), audio (*i.e.*, strong modality), and fusion recognition accuracy in multimodal learning, respectively. The green and purple dashed lines represent the best video and audio recognition accuracy in unimodal learning. (e) - (h) show the corresponding schematic illustration of the mutual information maximization result of the methods. Blue and red represent the excavated mutual information $I(\mathbf{X}^1;\mathbf{Z};\mathbf{Y})$ and $I(\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$ between modalities $\mathbf{X}^1, \mathbf{X}^2$, feature $\mathbf{Z}$ and labels $\mathbf{Y}$, and green represents unexcavated one. We reveal reserved complementary information ($\mathcal{P}(\mathbf{X}^1) = I(\mathbf{X}^1;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$ and $\mathcal{P}(\mathbf{X}^2) = I(\mathbf{X}^2;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$) and the upper limit of complementary information ($\tilde{\mathcal{P}}(\mathbf{X}^1) = I(\mathbf{X}^1;\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y})$ and $\tilde{\mathcal{P}}(\mathbf{X}^2) = I(\mathbf{X}^2;\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y})$) of respective modalities for an intuitive understanding of each method's characteristic.

advances, these approaches have not fully eradicated the impact of IML. Specifically, as visualized in Figures 2(a) to 2(c), the vanilla fusion method (*i.e.*, concatenation fusion without balance modulation) suffers from an extreme optimization imbalance, limiting overall performance improvement. While OGM-GE [2] (*i.e.*, the representative approach of the pace-adjustment-based methods) promotes optimization balance, both strong and lazy modalities often underperform relative to their unimodal counterparts. Similarly, although MMPareto [12] (a recent conflict-resolution-based approach) may even surpass unimodal learning in strong modalities, it tends to overlook the discriminative potential of lazy modalities, thereby obtaining sub-optimal overall performance.

Upon further investigation, we found that these methods essentially strive to maximize the mutual information between fused features and label semantics [16]. However, when confronted with IML, the heuristic balancing strategies they adopt do not align well with the objective of mutual information maximization, making it difficult to achieve a truly balanced fusion across modalities. To address these challenges, we present an analysis of IML from the perspective of information theory [17], establishing a theoretically complete framework for balanced multimodal learning. To be specific, we reveal a Multimodal Information Balance (MIB) theory, which characterizes IML as the imbalanced retention of complementary information (*i.e.*, $\mathcal{P}(\mathbf{X}^1) \gg \mathcal{P}(\mathbf{X}^2)$) during the process of maximizing mutual information between multimodal representations and label semantics, as illustrated in Figure 1. This analysis further explains why existing methods fall short of addressing imbalanced learning, as visually observed in Figures 2(e) to 2(g). Unfortunately, without any corrective measures, vanilla fusion method is directly affected by imbalanced optimization and struggles to preserve complementary information of the lazy modality (*i.e.*, $\mathcal{P}(\mathbf{X}^1) \gg \mathcal{P}(\mathbf{X}^2)$). Moreover, due to the unpredictability of information distribution, pace-adjustment-based methods, relying on empirical modulation, often fail to achieve sufficient mining of complementary information (*i.e.*, $\mathcal{P}(\mathbf{X}^1) < \tilde{\mathcal{P}}(\mathbf{X}^1)$ and $\mathcal{P}(\mathbf{X}^2) < \tilde{\mathcal{P}}(\mathbf{X}^2)$, where $\tilde{\mathcal{P}}(\mathbf{X}^1)$ and $\tilde{\mathcal{P}}(\mathbf{X}^2)$ denote the upper limits of complementary information for the respective modalities, as shown in Figure 2). Conflict-resolution-based methods address conflicts at the intersection of strong- and lazy-modality information, resulting in only facilitating the comprehensive extraction of the shared information between modalities (*i.e.*, $I(\mathbf{X}^1, \mathbf{X}^2;\mathbf{Z};\mathbf{Y})$). Consequently, although these methods could fully exploit the information in the strong modality, they neglect the complementary information of the lazy modality (*i.e.*, $\mathcal{P}(\mathbf{X}^2) \ll \tilde{\mathcal{P}}(\mathbf{X}^2)$), potentially disregarding task-beneficial discriminative information. (For necessary definitions, please refer to Section 3.2.)

To address this issue, we first develop an intuitive criterion based on MIB theory that guides balanced multimodal learning by adaptively balancing the preservation of complementary information across distinct modalities, thereby maximizing the mutual information between modalities and label semantics impartially and comprehensively. To implement this criterion, we propose a deep Information-Balanced Multimodal Learning (IBML) framework, which comprises two complementary modules: i) a Balance In-

formation Optimization (BIO) module that enforces the MIB criterion to facilitate the multimodal balance during joint optimization, and ii) a Task Complexity Modulation (TCM) module that enhances the balance of complementary information preservation by modulating modality-specific data complexity. Specifically, BIO directly derives a tractable lower bound for the MIB criterion to construct the MIB loss. Minimizing this loss enables BIO to adaptively balance the optimization of complementary information preservation across different modalities during multimodal fusion, thereby embracing the capability to exploit information within various modalities impartially. Inspired by the observed correlation between task complexity and complementary information preservation, our TCM dynamically adjusts the multimodal data to reduce task complexity discrepancies between strong and lazy modalities. By narrowing the discrepancies, TCM could indirectly enforce balanced preservation in line with the MIB criterion, ensuring comprehensive balance throughout the learning process. By integrating BIO and TCM, our IBML could effectively implement the MIB criterion at both optimization and data, achieving more comprehensive and balanced multimodal learning, as shown in Figures 2(d) and 2(h).

Our main contributions can be summarized as follows:

- We derive an information theory into Imbalanced Multimodal Learning (IML) understanding and propose a Multimodal Information Balance (MIB) criterion. Building on this, we present an Information-Balanced Multimodal Learning (IBML) framework that adaptively facilitates balanced multimodal learning.
- To directly enforce the MIB criterion for balanced complementary information retention, we propose a novel Balance Information Optimization (BIO) module that derives an optimizable lower bound for the MIB criterion, thereby ensuring balanced multimodal information fusion at the optimization side.
- To further promote balanced preservation of complementary information, a Task Complexity Modulation (TCM) module is presented to dynamically mitigate the modality-specific task complexity discrepancies at the data side, indirectly promoting the balance in accordance with the MIB criterion.
- Extensive experiments across three common multimodal task settings on eight datasets demonstrate the effectiveness of our proposed method, with IBML remarkably outperforming state-of-the-art approaches without bells and whistles.

2 Related Work

2.1 Imbalanced Multimodal Learning

Multimodal learning has attracted increasing attention from both academia and industry due to its ability to process multi-sensory data in real-world tasks, such as audio-visual recognition [18], [19] and image-text understanding [20], [21]. However, several recent studies [22], [23] have observed that joint multimodal training with exponentially more data does not always yield the expected performance gains over unimodal approaches, attributing this to imbalanced optimization across modalities, termed Imbalanced Multimodal Learning (IML). For instance, Wang *et al.* [10] demonstrate that models tend to favor a strong modality that dominates the optimization process, thereby neglecting the contributions of other modalities. Complementary analyses by Wu *et al.* [11], Huang *et al.* [24], and Ni *et al.* [25] further dissect IML from the perspectives of modal competition, greedy optimization, and model inherent properties. To mitigate IML, recent approaches can be broadly categorized into two groups: i) **Pace-adjustment-based methods** [2], [10], [26], [27], [28], [29] empirically analyze and adjust data augmentation, learning paces, optimization gradient, and loss functions to rebalance the optimization across modalities; In contrast, ii) **conflict-resolution-based methods** [12], [15], [30], [31], [32] focus on decoupling the dependencies across modalities and alleviating the inherent conflicts in optimization directions for the effective exploitation of modal information during joint multimodal learning. Despite their improvements, these solutions are predominantly heuristic and often fall short when addressing complex real-world tasks.

In this paper, we extend the investigation of IML by leveraging information theory to analyze the retention of modality-specific complementary information during mutual information maximization between multimodal features and label semantics. Different from prior works [11], [24], [31], [32] that focus solely on optimization dynamics, our approach not only attempts to analyze IML from a novel perspective of information theory but also establishes a criterion for achieving comprehensively balanced multimodal learning. This criterion informs our learning objectives and modulation schemes, offering a more intuitive and general approach compared to existing empirical modulation methods.

2.2 Information Theory in Multimodal Learning

Information theory [17] has found extensive application in multimodal learning for its ability to quantify the importance and differences of abstract information across different modalities. More specifically, concepts like entropy and information gain in information theory provide methods to model the quality of various feature distributions, providing firm foundations for multimodal feature selection and representation [33], [34], [35]. In addition, measures like joint entropy, Rényi relative entropy [36], and mutual information in information theory can be used to assess the disparities between two distributions. It drives them to serve as criteria to guide alignment across different modalities, achieving discriminative multi-view/modal clustering [37], [38], cross-modal learning [39], [40]. Moreover, the information bottleneck theory is applied in multimodal learning to suppress task-irrelevant or semantically redundant information in each modality, thereby obtaining more robust and consistent feature representations [41], [42]. Especially, maximizing the mutual information [43], or even the ternary mutual information [44], [45], across modalities is widely used as the objective for seeking optimal joint learning across modalities [46], [47] and agents [48] in general multimodal learning. However, achieving such ideal objectives is challenging due to the presence of imbalanced optimization during multimodal joint learning.

To tackle this challenge, this paper tries to explain imbalanced optimization from the perspective of imbal-

anced complementary information retention of respective modalities during mutual information maximization between multimodal features and label semantics. Based on this, we derive a novel information-balanced framework to endow each modality with an ideal information contribution, achieving balanced and comprehensive multimodal learning.

3 METHOD

To facilitate clear exposition, we first introduce the notations and formal definitions for imbalanced multimodal learning in Section 3.1. Next, drawing on information theory [17], we propose a Multimodal Information Balance (MIB) theory to interpret the imbalanced optimization problem and establish the corresponding MIB criterion in Section 3.2. Finally, in Section 3.3, we detail the proposed Information-Balanced Multimodal Learning (IBML) framework, which is designed to overcome the challenges of imbalanced optimization in multimodal joint learning.

3.1 Preliminaries and Problem Statement

For clarity, we begin by formally defining the imbalanced optimization problem in multimodal fusion to study the solution conveniently. Without loss of generality, consider a classification task with $M(M \geq 2)$ modalities. Let the multimodal dataset with C classes be denoted as $\mathcal{D} = \{(\{\mathbf{x}_i^m\}_{m=1}^M, y_i)\}_{i=1}^N$, where N is the number of labeled instances and $y_i \in \mathcal{Y} = \{1, 2, \cdots, C\}$. To perform multimodal recognition, we commonly employ M modality-specific encoders, *i.e.*, $\{\varphi_m(\Theta_m, \cdot)\}_{m=1}^M$, to obtain latent representations, where Θ_m represents the parameter set for the m-th encoder. Here, the encoding operation is defined as $\varphi_m(\mathbf{x}^m) : \mathcal{X}_m \mapsto \mathbb{R}^{1\times d_m}$, where $\mathcal{X}_m$ is the input space for the m-th modality, and d_m denotes its feature dimensionality. For simplicity, using vanilla concatenation fusion, the linear classifier applied to the fused representation is given by $g(\{\mathbf{x}_m\}_{m=1}^M) = [\varphi_1(\mathbf{x}_1^1); \cdots; \varphi_M(\mathbf{x}_i^M)]\mathbf{W} + \boldsymbol{b}$, where $g(\cdot) \in \mathbb{R}^{1\times C}$ produces the logits, $\mathbf{W} \in \mathbb{R}^{(\sum_{m=1}^M d_m)\times C}$ denotes the classifier parameters, and $\boldsymbol{b}$ is the bias term. The logit corresponding to the c-th class is denoted as $g(\cdot)_c$. Notably, the unimodal contributions can be decoupled as:

$$g(\{\mathbf{x}_i^m\}_{m=1}^M) = \sum_{m=1}^M g^m(\mathbf{x}_i^m) = \sum_{m=1}^M \varphi_m(\mathbf{x}_i^m)\mathbf{W}^m + \frac{\boldsymbol{b}}{M}, \tag{1}$$

where $\mathbf{W}^m \in \mathbb{R}^{d_m\times C}$. The model is typically optimized using gradient descent on the cross-entropy (CE) loss: $\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^N \log \sigma(g(\{\mathbf{x}_i^m\}_{m=1}^M))_{y_i}$, with $\sigma(\cdot)$ representing the $\mathrm{softmax}$ function. For i-th instance $\{\mathbf{x}_i^m\}_{m=1}^M$, the gradient with respect to the true label y_i is

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{CE}}}{\partial g^m(\mathbf{x}_i^m)_{y_i}} &= \frac{\partial \mathcal{L}_{\text{CE}}}{\partial g(\{\mathbf{x}_i^m\}_{m=1}^M)_{y_i}} \cdot \frac{\partial g(\{\mathbf{x}_i^m\}_{m=1}^M)_{y_i}}{\partial g^m(\mathbf{x}_i^m)_{y_i}} \\ &= \frac{\partial \mathcal{L}_{\text{CE}}}{\partial g(\{\mathbf{x}_i^m\}_{m=1}^M)_{y_i}} = \sigma(g(\{\mathbf{x}_i^m\}_{m=1}^M))_{y_i} - 1.\end{aligned} \tag{2}$$

As evident from Equation (2), when the fusion logits become over-confident and one modality contributes much more than another one on gradients, the optimization (gradients) of the lazy modality is suppressed since lagging behind in optimization, resulting in suboptimal performance. This phenomenon, termed imbalanced optimization problem [2], [8], often arises due to the inherent differences in modalities, wherein strong modalities dominate the optimization process while weaker (lazy) modalities receive insufficient updates. This issue is formally defined as follows:

Definition 1. ***Imbalanced Multimodal Learning (IML):*** *Modalities yielding over-confident predictions dominate the optimization process while others are under-optimized. That is, for some $k \neq j$, $g^k(\mathbf{x}_i^k)_{y_i} \gg g^j(\mathbf{x}_i^j)_{y_i}, k \neq j$ during training.*

Recent works [2], [8], [14], [49] have analyzed IML from an optimization perspective, suggesting that insufficient gradients for lazy modalities lead to suboptimal learning when a strong modality dominates. Thus, we aim to perform balanced multimodal learning to fully exploit information on multiple modalities, thus fully boosting learning tasks. While prior works [2] have attempted to alleviate IML from an optimization perspective using heuristic and empirical solutions, we find that these methods do not strike the ideal balance in the analysis of information theory, as shown in Figures 2 and 6. In contrast, our work systematically analyzes and tackles IML by leveraging the information theory [17] to achieve balanced multimodal learning. Based on this, we propose an information-balanced multimodal learning framework to restore balance among modalities, ensuring comprehensive and effective fusion.

3.2 Multimodal Information Balance

Different from existing gradient modulation methods [2], we start with information theory. As a foundational framework for data processing, information theory [17] has been widely used in deep learning with success, which is centered on the concept of entropy. For a discrete random variable x, the entropy is defined as $H(x) = -\sum_x p(x)\log p(x)$, where $p(x)$ is the probability distribution of x. Generally speaking, the Mutual Information (MI) between two discrete random variables (x, y) is $I(x; y) = H(x) - H(x|y)$, where $H(x|y) = \sum_{x,y} p(x, y)\log p(x|y)$ is the conditional entropy, and $p(x, y)$ is the joint distribution. Now, consider two modalities, denoted by $\mathbf{X}^1$ and $\mathbf{X}^2$, with the fusion variable represented as $\mathbf{Z}$. Let P_{X^1Y} denote the joint distribution over $\mathcal{X}_1 \times \mathcal{Y}$, with $(\mathbf{X}^1, \mathbf{Y}) \sim P_{X^1Y}$ being a pair of random variables. Note that other joint distributions can be defined similarly. In multimodal joint fusion, our objective is to fully exploit the data to maximize the mutual information $I(\mathbf{Z}; \mathbf{Y})$ between the fusion variable and the labels. However, solely maximizing $I(\mathbf{Z}; \mathbf{Y})$ does not guarantee the fair or balanced treatment between different modalities and may lead to one modality dominating the optimization process, thereby resulting in suboptimal results, as illustrated in Figures 1 and 2.

To achieve balanced multimodal learning, we propose a Multimodal Information Balance (MIB) criterion based on the above defect and information theory, which could guide us to design algorithms for multimodal fusion, thus getting out of the quagmire of IML. More specifically, multimodal learning aims to maximize mutual information between each modality and ground truths while ensuring the balance in complementary information retention. However, this can

be tricky in IML due to the discrepancy between different modalities. We further observe from Figure 1 that the IML phenomenon is reflected in the complementary information ($\mathcal{P}(\mathbf{X}^1)$ and $\mathcal{P}(\mathbf{X}^2)$) retained by modality fusion, *i.e.*, $\mathcal{P}(\mathbf{X}^1) = I(\mathbf{X}^1;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$ and $\mathcal{P}(\mathbf{X}^2) = I(\mathbf{X}^2;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$, where $I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$ represents the shared information. Natural *motivation* is to maintain a balance between private attributes to achieve the multimodal learning goal while maximizing shared information.

To monitor the multimodal balance, we introduce an indicator $\rho^{1,2}$ indicating the discrepancy ratio across different modalities. Intuitively, if $\rho^{1,2} > 1$, the optimization of the first modality is more dominant than that of the second modality, and vice versa. Using this indicator, we present the MIB criterion to adaptively and impartially increase the complementary information of respective modalities. Mathematically, the criterion can thus be formulated as:

$$\begin{aligned}\mathcal{J}_{\text{mib}} = \max_{\Theta_1,\Theta_2,\mathbf{W}} (1 - \sum_{m=1}^{2} \mathbb{I}_m) \cdot I(\mathbf{Z};\mathbf{Y}) \\ + \mathbb{I}_1 \cdot \mathcal{P}(\mathbf{X}^1) + \mathbb{I}_2 \cdot \mathcal{P}(\mathbf{X}^2) + I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y}),\end{aligned} \tag{3}$$

where $\mathbb{I}_1$ (or $\mathbb{I}_2$) equals 1 if $\rho^{1,2} < 1$ (or $\rho^{1,2} > 1$), and 0 otherwise. The first term in Equation (3) targets the ideal balanced optimization, while the remaining terms aim to maintain a balance in complementary information and maximize shared information when an imbalance occurs. Furthermore, since $\mathcal{P}(\mathbf{X}^m) = I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Z};\mathbf{Y})$, we can merge the components in Equation (3) as:

$$\max_{\Theta_1,\Theta_2,\mathbf{W}} (1 - \sum_{m=1}^{2} \mathbb{I}_m) \cdot I(\mathbf{Z};\mathbf{Y}) + \sum_{m=1}^{2} \mathbb{I}_m \cdot I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y}). \tag{4}$$

For M modalities, we can similarly establish the criterion of MIB. Unlike the case involving two modalities, the discrepancy ratio involving multiple modalities is no longer a comparative indicator, but a comprehensive indicator. We will discuss its definition in Section 3.3. Like Equation (4), we define a generic criterion of MIB for balanced multimodal learning as follows:

$$\begin{aligned}\mathcal{J}_{\text{mib}} = \max_{\{\Theta_m\}_{m=1}^M,\mathbf{W}} \sum_{m=1}^{M} \mathbb{I}_m \cdot I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y}) \\ + \max(1 - \sum_{m=1}^{M} \mathbb{I}_m, 0) \cdot I(\mathbf{Z};\mathbf{Y}),\end{aligned} \tag{5}$$

where $\mathbb{I}_m$ is 1 if the complementary information of the m-th modality is under-optimized, and 0 otherwise. Note that Equation (4) is a special case of Equation (5) for $M = 2$.

3.3 Information-Balanced Multimodal Learning

Based on the above analyses and definitions, we propose an Information-Balanced Multimodal Learning framework (IBML) for achieving balanced multimodal learning, as illustrated in Algorithm 1. IBML consists of two core modules, **Balance Information Optimization (BIO)** and **Task Complexity Modulation (TCM)**. The BIO is designed to directly enforce the MIB criterion to facilitate balanced optimization, while TCM modulates raw input data to adjust modality-specific task complexities, thereby avoiding the need for post-hoc gradient modulation and enhancing scalability. We now detail these components.

Algorithm 1 The training process of IBML

Input: The training data $\mathcal{D} = \{(\{\mathbf{x}_i^m\}_{m=1}^M, y_i)\}_{i=1}^N$, the modality-specific encoders $\{\varphi_m\}_{m=1}^M$, the classifier g, the maximal epoch N_e;

```
Initialize the encoders and classifier;
for e = 1, 2, ···, N_e do
    for each mini-batch D' ⊆ D do
        if Conduct modulation then
            Calculate optimization strengths {ρ̃^m}_{m=1}^M based on the recorded predictions;
            // Task Complexity Modulation
            Inject noise into input data by Equation (16);
            // Balance Information Optimization
            Calculate L_mib by Equation (12);
        else
            Calculate L_CE for fusion logits;
        end if
        Record all decoupled and fusion predictions;
        Update all learnable parameters by SGD;
    end for
end for
```

Output: The optimized parameters.

3.3.1 Balance Information Optimization

In multimodal models, the over-dominance of a strong modality could lead to the under-optimization of a weaker (lazy) modality, thereby limiting overall performance. To overcome this issue, we enforce the MIB criterion in Equation (5) via Balance Information Optimization (BIO), boosting balanced multimodal learning. The criterion is decomposed into two objectives:

$$\max_{\{\Theta_m\}_{m=1}^M,\mathbf{W}} I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y}) \quad \text{and} \quad \max_{\{\Theta_m\}_{m=1}^M,\mathbf{W}} I(\mathbf{Z};\mathbf{Y}). \tag{6}$$

However, directly estimating and maximizing $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ is challenging due to interactions among three variables. Instead, we derive tractable lower bounds, which is a widely adopted strategy in mutual information optimization [16], [50]. We first decompose $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ as:

$$\begin{aligned}I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y}) =& I(\mathbf{Z};\mathbf{Y}) + I(\mathbf{X}^m;\mathbf{Y}) + H(\mathbf{X}^m,\mathbf{Z},\mathbf{Y}) \\ & - H(\mathbf{X}^m) - H(\mathbf{Z}) + I(\mathbf{X}^m;\mathbf{Z}) - H(\mathbf{Y}) \\ =& I(\mathbf{Z};\mathbf{Y}) + I(\mathbf{X}^m;\mathbf{Y}) - H(\mathbf{Y}) \\ & + H(\mathbf{X}^m,\mathbf{Z},\mathbf{Y}) - H(\mathbf{X}^m,\mathbf{Z}).\end{aligned} \tag{7}$$

Since $H(\mathbf{X}^m,\mathbf{Z},\mathbf{Y}) - H(\mathbf{X}^m,\mathbf{Z}) \geq 0$, $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ obviously has a strict lower bound of $I(\mathbf{Z};\mathbf{Y}) + I(\mathbf{X}^m;\mathbf{Y}) - H(\mathbf{Y})$, where $H(\mathbf{Y})$ is a constant and can be omitted during optimization in practice. Thus, we can achieve the objective in Equation (4) by maximizing its lower bound alternatively:

$$\begin{aligned}\mathcal{J}_{\text{mib}} \equiv \max_{\{\Theta_m\}_{m=1}^M,\mathbf{W}} \sum_{m=1}^{M} \mathbb{I}_m \cdot \big(I(\mathbf{Z};\mathbf{Y}) + I(\mathbf{X}^m;\mathbf{Y})\big) \\ + \max(1 - \sum_{m=1}^{M} \mathbb{I}_m, 0) \cdot I(\mathbf{Z};\mathbf{Y}).\end{aligned} \tag{8}$$

From this, the final objective is decomposed into maximizing $I(\mathbf{X}^m;\mathbf{Y})$ and $I(\mathbf{Z};\mathbf{Y})$, simultaneously. To achieve this, we can maximize their variational lower bounds. For $I(\mathbf{X}^m;\mathbf{Y})$, it holds:

$$I(\mathbf{X}^m;\mathbf{Y}) = \mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}\left[\log\frac{P(y|\mathbf{x})}{P(y)}\right] \geq \mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}\left[\log\frac{Q(\mathbf{x}^m,y)}{P(\mathbf{x}^m)P(y)}\right],$$

where P is the marginal distribution, Q is the variational distribution, and $\mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}$ abbreviates the expectation taken over samples drawn from the true joint data distribution, *i.e.*, $\mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})\sim P_{X^mY}}$, where P_{X^mY} represents the underlying joint distribution of m-th modality inputs and their corresponding labels. To optimize this lower bound, we apply reparameterization [51] for $Q(\mathbf{x}^m,y)$, *i.e.*,

$$Q(\mathbf{x}^m,y) = \frac{P(\mathbf{x}^m)P(y)}{\mathbb{E}_{y'\sim P_Y}\left[\exp\left(g^m(\mathbf{x}^m)_{y'}\right)\right]}\exp\left(g^m(\mathbf{x}^m)_y\right), \quad (9)$$

where P_Y denotes the label distribution used when normalizing over all possible classes in the variational formulation. If the training data is sufficiently large to adequately reflect the data distribution, we could get approximation of $\mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}\left[\log\frac{Q(\mathbf{x}^m,y)}{P(\mathbf{x}^m)P(y)}\right]$ by conducting parameterization [51] to obtain a optimizable lower bound, *i.e.*,

$$I(\mathbf{X}^m;\mathbf{Y}) \geq \mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}\left[\log\sigma(g^m(\mathbf{x}^m))_y\right] + \mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}\left[\log C\right], \quad (10)$$

where $\sigma(\cdot)$ is the $\mathrm{softmax}$ function and C is the number of classes. Similarly, the optimizable lower bound of $I(\mathbf{Z},\mathbf{Y})$ can be:

$$I(\mathbf{Z};\mathbf{Y}) \geq \mathbb{E}_{(\mathbf{Z},\mathbf{Y})}\left[\log\sigma(\mathbf{z})_y\right] + \mathbb{E}_{(\mathbf{Z},\mathbf{Y})}\left[\log C\right], \quad (11)$$

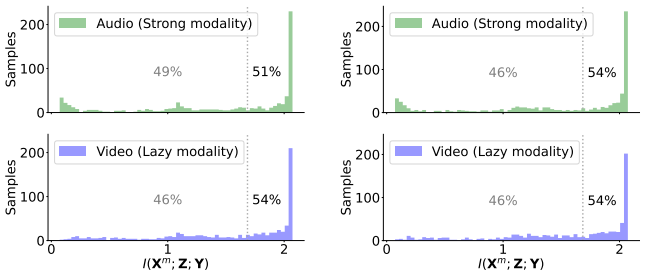
where $\mathbf{z} = g(\{\mathbf{x}^m\}_{m=1}^M)$ is the fusion logits of M modalities and $\mathbb{E}_{(\mathbf{Z},\mathbf{Y})\sim P_{ZY}}$ denotes the expectation over the joint distribution of the fused representation $\mathbf{Z}$ and label $\mathbf{Y}$, analogous to the definition of P_{X^mY} for unimodal inputs. The proofs of Equations (9) to (11) can be found in the Appendix. Then, we put them together and obtain the MIB loss, *i.e.*,

$$\begin{aligned}\mathcal{L}_{\text{mib}} = &-\hat{\mathbb{I}}\cdot\mathbb{E}_{(\mathbf{Z},\mathbf{Y})}\left[\log\sigma(\mathbf{z})_y\right] \\ &-\sum_{m=1}^{M}\mathbb{I}_m\cdot\Big(\mathbb{E}_{(\mathbf{Z},\mathbf{Y})}\left[\log\sigma(\mathbf{z})_y\right] \\ &+\mathbb{E}_{(\mathbf{X}^m,\mathbf{Y})}\left[\log\sigma(g^m(\mathbf{x}^m))_y\right]\Big) + \text{Const},\end{aligned} \quad (12)$$

where $\hat{\mathbb{I}} = \max(1-\sum_{m=1}^M\mathbb{I}_m, 0)$ and Const is a constant term related to $\log C$. For simplicity, the constant term can be dropped since it does not produce gradients in practical multimodal training.

To minimize $\mathcal{L}_{\text{mib}}$, another important thing is to monitor the optimization discrepancies of M modalities to determine $\{\mathbb{I}_m\}_{m=1}^M$ and then selectively optimize $\mathcal{L}_{\text{mib}}$ for balance. Inspired by [2], given a mini-batch of $\mathcal{D}' = \{(\{\mathbf{x}_k^m\}_{m=1}^M, y_k)\}_{k=1}^K$, we define $\{\rho^m\}_{m=1}^M$ for all modalities to indicate the optimization strength, where ρ^m is calculated by:

$$\rho^m = \sum_{k=1}^{K}\sigma(g^m(\mathbf{x}_k^m))_{y_k}. \quad (13)$$



(a) *w/o* TCM (b) Full

Fig. 3: The distribution comparison on the $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ of samples before (*i.e.*, *w/o* TCM) and after adopting the TCM (*i.e.*, Full) on the CREMA-D dataset. The green subfigures represent the strong modality, and the blue ones represent the lazy modality. Each subfigure is divided into two parts, *i.e.*, high $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ and low $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$, and the black and gray percentages represent the sample proportion of the two parts, respectively.

By monitoring with $\{\rho^m\}_{m=1}^M$, we can dynamically perceive the contribution discrepancy across all modalities. Finally, we set $\mathbb{I}_m$ as 1 if $\rho^m < \frac{1}{M}\sum_{m=1}^M\rho^m$, and 0 otherwise. After all of the above, one can see that minimizing Equation (12) is equivalent to achieving the objective of the criterion as shown in Equation (5).

3.3.2 Task Complexity Modulation

Although the proposed BIO modulates the optimization from the direct model prediction, we observe that BIO alone cannot achieve the optimal balance for the MIB criterion. Specifically, we estimate the core component $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ in the MIB criterion (*i.e.*, Equation (5)) using the computable lower bound. As shown in Figure 3(a), the proportion of the samples with high $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ remains limited for both lazy and strong modalities, reaching only 51% and 54%, respectively. To address this limitation, we propose Task Complexity Modulation (TCM) to complement BIO based on the information theory and achieve balanced modulation across multimodal inputs.

Recent studies [52], [53] reveal that the complexity of a given task $\mathcal{T}$ can be quantified by its entropy, which is equal to the conditional entropy $H(\mathcal{T}|\mathbf{X}^m)$ when conditioned on data $\mathbf{X}^m$. To be specific, now considering a common single-label classification scenario, the "*uncertainty*" or "*complexity*" of observed labels $\mathbf{Y}$ given the data $\mathbf{X}^m$, reflecting the task complexity conditioned on $\mathbf{X}^m$, and can be formulated as:

$$H(\mathcal{T}|\mathbf{X}^m) = H(\mathbf{Y}|\mathbf{X}^m) = H(\mathbf{Y}) - I(\mathbf{X}^m;\mathbf{Y}). \quad (14)$$

When label semantic space is specified, a lower $H(\mathbf{Y}|\mathbf{X}^m)$ implies that $\mathbf{X}^m$ provides a stronger explanation for $\mathbf{Y}$, corresponding to higher mutual information $I(\mathbf{X}^m;\mathbf{Y})$, and vice versa. Consequently, our natural motivation is that *balancing multimodal learning can be achieved by modulating task complexity*. To achieve this, we explore the relationship between data $\mathbf{X}^m$, task complexity $H(\mathcal{T}|\mathbf{X}^m)$, and the mutual information $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$, as presented in Lemma 1.

Lemma 1. *Given a specific task $\mathcal{T}$, the task complexity conditioned on data $\mathbf{X}^m$, i.e., $H(\mathcal{T}|\mathbf{X}^m)$, and its corresponding mutual information of $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ with the fusion of M modalities*

$\{\mathbf{X}^1, \cdots, \mathbf{X}^M\}$, *when injecting data perturbation to the data, i.e., there is a Markov chain [54]* $\mathbf{X}^m \rightarrow \tilde{\mathbf{X}}^m \rightarrow \mathbf{Y}$, *the complexity* $H(\mathcal{T}|\tilde{\mathbf{X}}^m)$ *increases and we have:*

$$H(\mathcal{T}|\tilde{\mathbf{X}}^m) \geq H(\mathcal{T}|\mathbf{X}^m) \Rightarrow I(\mathbf{X}^m; \mathbf{Z}; \mathbf{Y}) \geq I(\tilde{\mathbf{X}}^m; \tilde{\mathbf{Z}}; \mathbf{Y}).$$

Lemma 2. *Injecting independent gaussian noise with variance* $\sigma_\epsilon > 0$ *into the training data can increase the complexity of* $\mathcal{T}$*, i.e.,* $H(\mathcal{T}|\mathbf{X} + \boldsymbol{\epsilon}) \geq H(\mathcal{T}|\mathbf{X})$*, where* $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon)$*.*

Proof. The proofs of Lemmas 1 and 2 are provided in the Appendix. □

Based on Lemma 1, one can see that directly modulating the input data by forming a Markov chain can directly increase task complexity, thereby reducing $I(\mathbf{X}^m; \mathbf{Z}; \mathbf{Y})$, which reflects the ultimate optimization goal of MIB criterion in Section 3.2. To form the Markov chain, inspired by studies on entropy change [52], [53], we introduce noise into the data $\mathbf{X}^m$ to controllably increase the complexity of the task $\mathcal{T}$, as mentioned in Lemma 2. Finally, similar to BIO, we selectively introduce data noise to increase task complexity. More specifically, given a mini-batch of $\mathcal{D}' = \{(\{\mathbf{x}_k^m\}_{m=1}^M, y_k)\}_{k=1}^K$, we modulate the task complexities for all modalities based on the record optimization strengths $(\{\tilde{\rho}^m\}_{m=1}^M)$, where $\tilde{\rho}^m = \sum_{k=1}^K \sigma(\tilde{g}^m(\mathbf{x}_k^m))_{y_k}$ and $\tilde{g}^m(\mathbf{x}_k^m)$ is the logits of sample $\mathbf{x}_k^m$ recorded at the previous training epoch. The modulation process for all input samples is:

$$\hat{\mathbf{x}}_k^m = \mathbf{x}_k^m + \boldsymbol{\epsilon}^m, \forall \mathbf{x}_k^m \in \mathcal{D}', \text{ if } \tilde{\rho}^m > \frac{1}{M}\sum_{m=1}^M \tilde{\rho}^m, \quad (15)$$

where $\boldsymbol{\epsilon}^m \sim \mathcal{N}(0, \mathrm{var}(\{\mathbf{x}_k^m\}_{k=1}^K))$ is the independent gaussian noise added to the m-th modal samples. In practice, although increasing the complexity of the task through Equation (15) is conducive to balanced optimization, injecting noise into all samples of a task modality can easily destroy its original information pattern, resulting in suboptimal performance. For this reason, we recommend injecting noise in a random probability manner only to the samples that the training logits predict correctly. This can prevent under-optimized samples from being affected by noise while increasing the task complexity of strong modalities, thereby achieving a trade-off between task complexity and task performance. For $\forall \mathbf{x}_k^m \in \mathcal{D}'$ and $y_k = \arg\max \sigma(g^m(\mathbf{x}_k^m))$, the modulation process is re-described as:

$$\hat{\mathbf{x}}_k^m = \mathbf{x}_k^m + \boldsymbol{\epsilon}^m, \text{ if } \tilde{\rho}^m > \frac{1}{M}\sum_{m=1}^M \tilde{\rho}^m \text{ and } p_k^m < p^m, \quad (16)$$

where $p_k^m \in [0, 1]$ is a random probability value and $p^m = \min(\frac{M\tilde{\rho}^m}{\sum_{m=1}^M \tilde{\rho}^m}, \lambda)/\lambda$, which applies a certain probability of injecting noise to over-confident samples based on the degree of optimization imbalance, increasing the overall complexity of the strong modality. λ serves as a trade-off noise scaling factor, controlling the sensitivity to imbalance during the noise injection process.

4 Experiments

In this section, we conduct comprehensive experiments on eight widely-used multimodal datasets spanning three tasks (*i.e.*, audio-visual recognition, image-text classification, and 2D-3D recognition) to evaluate the effectiveness and superiority of our proposed IBML.

TABLE 1: Brief statistics of the used datasets. 'AV' means the task of audio-visual recognition, 'VL' means the task of image-text classification, and '23D' means the task of 2D-3D recognition.

Datasets	Task	Categories	Training	Validation	Testing
CREMA-D	AV	6	6,697	744	744
VAE	AV	28	3,312	402	401
AVSBench	AV	23	3,452	740	740
VGGSound50	AV	50	25,954	-	2,499
MVSA	VL	3	1,555	519	518
FOOD101	VL	101	62,971	22,715	5,000
3D MNIST	23D	10	5,000	-	1,000
ModelNet40	23D	40	9,840	-	3,991

4.1 Datasets

We evaluate our method on eight widely-used multimodal datasets across three tasks, including audio-visual recognition, image-text classification, and 2D-3D recognition. Table 1 provides brief statistics for these datasets, and further details can be found in the Appendix.

4.2 Experiment Settings

To ensure fair comparisons, we employ unified modality-specific backbones for all methods under the same task throughout our experiments. Specifically:

- **Audio-Visual Recognition:** Following [2], for the video modality, we randomly select three frames from each video per iteration as visual inputs, which are then fed into a ResNet18-based network for feature encoding under this task. For the audio modality, we modify the ResNet18 network by changing its input channel from 3 to 1, thus enabling it to process one-dimensional data.
- **Image-Text Classification:** In this task, a ResNet18-based network encodes input images, while textual data is processed using a Glove embedding layer [55] followed by a Bi-GRU network.
- **2D-3D Recognition:** For 2D image modality, a simple CNN encodes the low-resolution images in 3D MNIST, and we use a ResNet18-based multi-channel network to encode multi-view images in ModelNet40, as in [56]. Additionally, we adopt the widely-used DGCNN [57] to encode 3D point-cloud objects.

For a fair comparison, all models are trained from scratch (except for the Glove embedding layer) using the SGD optimizer with a momentum of 0.9 and a weight decay of 1×10^{-4}. λ is set to 10 for appropriate modulation of task complexity in TCM. In the early stages of training, we employ a *warmup* strategy with a larger learning rate for our for IBML to facilitate rapid initial convergence. Notably, unlike existing methods (*e.g.*, OGM-GE [2], DGL [32], and InfoReg [29]), our method does not require any additional hyperparameter tuning and all experiments are performed on a single GeForce RTX3090 24GB GPU.

TABLE 2: Test accuracies (%) of eight baselines on four audio-visual recognition benchmark datasets. The results with the form of 'mean ± std' are reported over three random runs and the best results are boldfaced.

Methods	Test	CREMA-D	AVE	AVSBench	VGGSound50
Summation	Fusion	61.07 ± 1.30	66.33 ± 0.85	87.43 ± 0.90	53.73 ± 0.48
	Audio	53.15 ± 2.61	53.73 ± 1.42	76.67 ± 0.80	40.55 ± 0.26
	Video	22.18 ± 2.39	19.32 ± 1.73	25.36 ± 0.39	19.04 ± 0.82
Concatenation	Fusion	63.26 ± 0.91	66.42 ± 0.54	87.34 ± 0.45	53.82 ± 0.63
	Audio	54.75 ± 0.39	53.48 ± 1.81	77.66 ± 1.62	40.78 ± 0.27
	Video	24.51 ± 1.47	18.33 ± 0.59	22.89 ± 2.33	19.33 ± 0.35
OGM-GE [2] (CVPR'22)	Fusion	67.16 ± 1.23	65.67 ± 1.02	87.50 ± 0.67	55.96 ± 0.21
	Audio	50.13 ± 0.87	50.66 ± 1.19	74.41 ± 2.02	38.64 ± 0.14
	Video	29.93 ± 0.93	19.82 ± 1.73	26.37 ± 2.52	22.21 ± 0.18
PMR [8] (CVPR'23)	Fusion	65.16 ± 1.08	66.09 ± 0.47	89.23 ± 0.17	53.18 ± 0.74
	Audio	53.27 ± 0.64	51.24 ± 2.24	76.98 ± 1.39	37.32 ± 0.96
	Video	32.75 ± 0.83	12.03 ± 0.59	30.72 ± 2.82	19.91 ± 1.28
MMCosine [49] (ICASSP'23)	Fusion	63.73 ± 0.42	67.49 ± 0.47	89.19 ± 0.69	54.15 ± 0.33
	Audio	55.29 ± 1.58	55.72 ± 0.20	81.89 ± 0.80	43.50 ± 0.49
	Video	31.99 ± 2.50	20.98 ± 0.31	35.59 ± 3.89	25.65 ± 2.71
AGM [14] (ICCV'23)	Fusion	65.68 ± 0.73	68.32 ± 0.94	87.93 ± 0.88	53.94 ± 0.51
	Audio	57.66 ± 0.48	54.23 ± 2.54	79.28 ± 0.90	42.99 ± 0.23
	Video	26.84 ± 0.85	17.49 ± 1.75	24.68 ± 0.92	20.69 ± 0.64
MMPareto [12] (ICML'24)	Fusion	69.53 ± 1.20	69.73 ± 0.62	91.44 ± 0.93	59.06 ± 0.13
	Audio	61.02 ± 1.43	61.36 ± 1.12	81.80 ± 0.95	45.02 ± 0.09
	Video	46.64 ± 2.01	33.17 ± 1.12	54.05 ± 0.40	36.05 ± 0.69
MLA [15] (CVPR'24)	Fusion	74.42 ± 0.89	68.74 ± 0.47	84.59 ± 1.34	58.21 ± 0.77
	Audio	59.14 ± 0.88	64.10 ± 0.47	75.81 ± 2.65	46.91 ± 0.46
	Video	65.05 ± 0.29	32.18 ± 1.24	40.99 ± 0.17	35.16 ± 0.92
DGL [32] (ICCV'25)	Fusion	79.40 ± 1.59	68.82 ± 2.34	91.98 ± 0.33	57.40 ± 0.22
	Audio	61.88 ± 2.26	63.60 ± 1.52	83.97 ± 1.64	48.39 ± 0.57
	Video	72.96 ± 2.34	24.87 ± 1.33	63.54 ± 0.77	39.44 ± 0.47
ARL [31] (ICCV'25)	Fusion	78.87 ± 0.49	69.87 ± 0.44	91.94 ± 0.17	60.19 ± 0.79
	Audio	60.24 ± 1.32	62.50 ± 1.61	83.29 ± 0.29	46.31 ± 0.45
	Video	69.19 ± 1.49	37.50 ± 1.39	56.23 ± 1.12	38.91 ± 0.15
InfoReg [29] (CVPR'25)	Fusion	74.77 ± 0.77	69.89 ± 0.19	91.35 ± 0.51	59.46 ± 0.32
	Audio	59.13 ± 2.72	63.25 ± 0.97	81.80 ± 0.51	46.42 ± 0.36
	Video	61.00 ± 0.62	31.44 ± 0.93	52.88 ± 0.55	35.43 ± 0.72
Our method	Fusion	$\mathbf{80.78 \pm 0.77}$	$\mathbf{70.15 \pm 1.07}$	$\mathbf{92.34 \pm 0.39}$	$\mathbf{60.54 \pm 0.57}$
	Audio	57.39 ± 0.48	59.95 ± 0.71	78.69 ± 1.12	43.74 ± 0.53
	Video	71.06 ± 0.73	35.57 ± 0.93	59.64 ± 1.12	39.83 ± 0.21

4.3 Comparisons on Multimodal Tasks

To verify the effectiveness and superiority of our IBML, we compare it with eleven baseline methods: two **vanilla fusion methods** (*i.e.*, Summation and Concatenation), four **pace-adjustment-based methods** (*i.e.*, OGM-GE (CVPR'22) [2], PMR (CVPR'23) [8], MMCosine (ICASSP'23) [49], AGM (ICCV'23) [14], InfoReg (CVPR'25) [29]), and four **conflict-resolution-based methods** (*i.e.*, MMPareto (ICML'24) [12], MLA (CVPR'24) [15], DGL (ICCV'25) [32], ARL (ICCV'25) [31]). Table 2 reports recognition accuracy for audio-visual recognition, while Table 3 presents results for image-text classification and 2D-3D recognition tasks. Additionally, Table 4 includes a comparison of the modal discrepancy ratio (*i.e.*, $\rho^i/\rho^j, i \neq j$) across the three tasks, which reflects the degree of optimization balance [2].

From our experimental results, we could draw the following observations and conclusions:

- **Impact of Imbalanced Optimization:** Imbalanced optimization remarkably degrades performance across these three tasks. In extreme cases (*e.g.*, CREMA-D and MVSA), the lazy modalities perform nearly at chance levels.
- **Strong Modality Preservation:** For strong modalities, pace-adjustment-based methods (*e.g.*, OGM-GE, PMR, MMCosine, and InfoReg) often underperform compared to vanilla fusion. In contrast, our IBML preserves the unimodal performance of strong modalities, demonstrating that their complexity modulation does not hinder the optimization of dominant modalities.
- **Lazy Modality Improvement:** Our IBML achieves notable performance gains for the lazy modality. For example, on the CREMA-D dataset, IBML boosts the recognition performance of the lazy modality by up to 200% compared with MMPareto, illustrating its capability to fully exploit discriminative information from lazy modalities and embrace balanced optimization.
- **Overall Superiority:** Across all tasks, our IBML outperforms existing multimodal balanced learning baselines,

TABLE 3: Test accuracies (%) of baselines on image-text classification (MVSA and FOOD101) and 2D-3D recognition (3D MNIST and ModelNet40) benchmark datasets. The results in the form of 'mean ± std' are reported over three random runs and the best results are boldfaced.

Methods	Test	MVSA	FOOD101	3D MNIST	ModelNet40
Summation	Fusion	72.19 ± 0.92	87.58 ± 0.04	96.80 ± 0.50	91.33 ± 0.25
	Image/2D	28.45 ± 0.36	29.42 ± 0.16	32.00 ± 10.1	87.20 ± 0.26
	Text/3D	65.64 ± 0.55	82.40 ± 0.09	96.70 ± 0.60	55.63 ± 1.81
Concatenation	Fusion	71.16 ± 0.55	87.50 ± 0.10	96.80 ± 0.30	91.32 ± 0.35
	Image/2D	29.67 ± 2.16	30.11 ± 0.38	87.12 ± 0.41	87.12 ± 0.26
	Text/3D	61.72 ± 0.55	87.50 ± 0.10	50.19 ± 1.47	50.19 ± 0.73
OGM-GE [2] (CVPR'22)	Fusion	70.84 ± 1.05	87.49 ± 0.07	97.30 ± 0.22	91.10 ± 0.36
	Image/2D	29.61 ± 0.96	30.39 ± 0.24	62.53 ± 8.57	85.67 ± 0.91
	Text/3D	60.18 ± 1.77	81.66 ± 0.13	96.67 ± 0.47	53.63 ± 1.99
PMR [8] (CVPR'23)	Fusion	74.62 ± 0.59	80.25 ± 0.59	96.97 ± 0.30	89.50 ± 0.92
	Image/2D	51.97 ± 0.89	31.21 ± 1.97	77.47 ± 6.40	81.27 ± 2.11
	Text/3D	75.25 ± 0.28	70.96 ± 0.16	93.27 ± 1.22	55.17 ± 4.33
MMCosine [49] (ICASSP'23)	Fusion	66.41 ± 0.51	88.73 ± 0.12	97.23 ± 0.12	91.17 ± 0.68
	Image/2D	37.76 ± 2.32	43.59 ± 0.16	74.00 ± 7.73	88.80 ± 1.04
	Text/3D	54.66 ± 1.00	82.89 ± 0.08	97.07 ± 0.06	76.50 ± 4.50
AGM [14] (ICCV'23)	Fusion	67.63 ± 0.51	87.18 ± 0.15	97.03 ± 0.21	90.86 ± 0.28
	Image/2D	30.12 ± 1.54	30.30 ± 0.48	80.43 ± 4.05	80.39 ± 0.56
	Text/3D	42.83 ± 1.03	79.57 ± 0.16	95.86 ± 0.15	78.25 ± 0.22
MMPareto [12] (ICML'24)	Fusion	72.04 ± 1.18	89.24 ± 0.12	98.23 ± 0.32	91.60 ± 0.24
	Image/2D	50.05 ± 2.58	49.60 ± 0.04	97.07 ± 0.35	89.51 ± 0.49
	Text/3D	74.63 ± 1.10	84.06 ± 0.05	96.87 ± 0.68	86.47 ± 0.54
MLA [15] (CVPR'24)	Fusion	61.87 ± 1.58	89.30 ± 0.43	98.16 ± 0.35	91.72 ± 1.25
	Image/2D	51.57 ± 1.76	51.70 ± 2.36	97.50 ± 0.26	90.27 ± 0.92
	Text/3D	68.80 ± 0.53	83.53 ± 0.06	96.37 ± 0.51	88.18 ± 0.49
DGL [32] (ICCV'25)	Fusion	69.62 ± 1.37	88.24 ± 0.01	97.10 ± 0.37	91.89 ± 0.40
	Image/2D	50.80 ± 0.78	56.51 ± 0.22	94.37 ± 3.94	89.87 ± 0.30
	Text/3D	51.32 ± 0.09	67.51 ± 0.76	96.57 ± 0.49	86.86 ± 0.58
ARL [31] (ICCV'25)	Fusion	70.26 ± 1.11	88.39 ± 0.20	97.99 ± 0.21	86.76 ± 0.15
	Image/2D	51.25 ± 2.65	56.32 ± 0.10	97.26 ± 0.12	91.88 ± 0.46
	Text/3D	51.32 ± 0.09	67.56 ± 0.84	97.46 ± 0.18	90.08 ± 0.25
InfoReg [29] (CVPR'25)	Fusion	74.05 ± 0.79	89.20 ± 0.15	97.93 ± 0.70	89.72 ± 1.11
	Image/2D	50.67 ± 1.10	84.26 ± 0.03	97.67 ± 0.40	86.10 ± 1.46
	Text/3D	74.57 ± 1.35	40.37 ± 0.24	94.43 ± 2.54	82.32 ± 1.07
Our method	Fusion	$\mathbf{75.21 \pm 0.64}$	$\mathbf{90.67 \pm 0.11}$	$\mathbf{98.63 \pm 0.06}$	$\mathbf{92.52 \pm 0.60}$
	Image/2D	51.96 ± 0.73	55.73 ± 0.46	97.60 ± 0.36	88.32 ± 0.24
	Text/3D	75.27 ± 0.24	84.02 ± 0.05	96.43 ± 1.16	87.77 ± 0.52

TABLE 4: Average discrepancy ratio during training on three datasets across different tasks. The results with the form of 'mean ± std' are reported over three random runs and the best results (*i.e.*, close to 1) are boldfaced. The second-best results are underlined.

Methods	CREMA-D	MVSA	3D MNIST
Concatenation	4.33 ± 0.33	1.83 ± 0.34	8.44 ± 0.20
OGM-GE [2]	2.86 ± 0.09	1.43 ± 0.02	2.24 ± 0.17
PMR [8]	2.98 ± 0.12	$\mathbf{0.93 \pm 0.40}$	1.30 ± 0.02
MLA [9]	3.74 ± 1.47	0.62 ± 0.11	1.44 ± 0.13
InfoReg [29]	$\underline{1.61 \pm 0.02}$	1.25 ± 0.01	$\underline{0.91 \pm 0.01}$
Our method[1]	$\mathbf{1.55 \pm 0.00}$	$\underline{1.24 \pm 0.01}$	$\mathbf{1.01 \pm 0.00}$

validating its effectiveness in mitigating imbalanced optimization and promoting multimodal learning.

1. The standard deviations on datasets CREMA-D and 3D MNIST are 0.002 and 0.00003, respectively. To preserve consistent and appropriate significant figures, they are uniformly reported as 0.00.

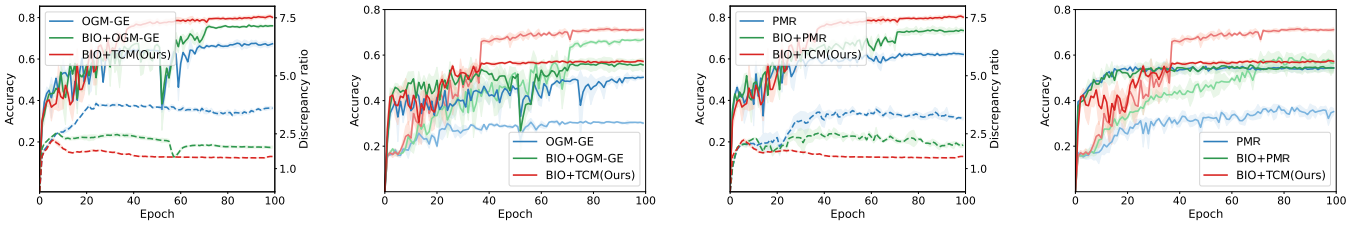
In summary, these experimental results clearly demonstrate that IBML exhibits strong stability and superior performance in multimodal learning, attributed to the synergistic efficacy of the proposed BIO and TCM. In the following sections, we give an in-depth analysis to explore the behavior and contribution of each proposed component in IBML.

4.4 Ablation Study

We conduct ablation studies on the CREMA-D dataset to investigate the individual contributions of the BIO module, TCM, and the warmup mechanism. In these experiments, we selectively remove one or two components and compare the modified variant against both the full IBML framework and the vanilla concatenation baseline. As shown in Table 5, we could draw the following observations: 1) Removing any single component exacerbates the imbalance in multimodal learning, resulting in a performance drop. 2) The inclusion of the *warmup* mechanism substantially enhances performance, indicating that rapid initial convergence is

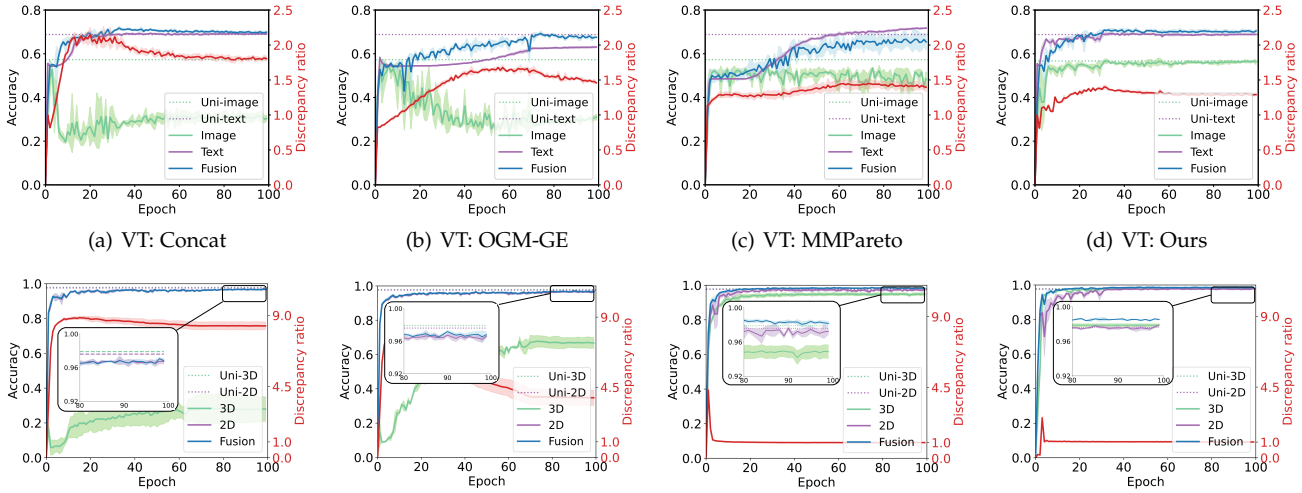
TABLE 5: Ablation experiments evaluated on the CREMA-D dataset.

No.	BIO	TCM	Warmup	Audio (↑)	Video (↑)	Fusion (↑)	Discrepancy Ratio (↓)
#1	✓	✓	✓	57.39 ± 0.48	$\mathbf{71.06 \pm 0.73}$	$\mathbf{80.78 \pm 0.77}$	$\mathbf{1.55 \pm 0.00}$
#2	✓	✗	✓	$\mathbf{59.36 \pm 0.06}$	70.07 ± 0.83	80.11 ± 0.66	1.64 ± 0.02
#3	✓	✓	✗	56.18 ± 0.00	66.04 ± 0.82	77.51 ± 0.32	1.87 ± 0.13
#4	✓	✗	✗	56.50 ± 0.81	66.13 ± 1.16	75.81 ± 0.48	2.00 ± 0.04
#5	✗	✓	✗	55.78 ± 1.81	30.15 ± 2.71	66.26 ± 0.76	3.21 ± 0.09
#6	Vanilla Concatenation			54.75 ± 0.39	24.51 ± 1.47	63.26 ± 0.91	4.19 ± 0.29



(a) Overall: OGM-GE (b) Specific: OGM-GE (c) Overall: PMR (d) Specific: PMR

Fig. 4: The modal discrepancy ratio and recognition accuracy of the OGM-GE [2], PMR [8], corresponding constructed baselines (*i.e.*, BIO+OGM-GE and BIO+PMR), and our IBML (*i.e.*, BIO+TCM) on the CREMA-D dataset [7]. (a) and (c) show the overall performances (*i.e.*, modal discrepancy ratio and fusion recognition accuracy) during the learning process. Solid lines represent the recognition accuracy, and the dashed ones represent the discrepancy ratio across modalities. (b) and (d) show the specific performances (*i.e.*, recognition accuracy of the strong and lazy modalities). Dark colors represent strong modalities, and corresponding light ones represent lazy modalities.



(a) VT: Concat (b) VT: OGM-GE (c) VT: MMPareto (d) VT: Ours

(e) 23D: Concat (f) 23D: OGM-GE (g) 23D: MMPareto (h) 23D: Ours

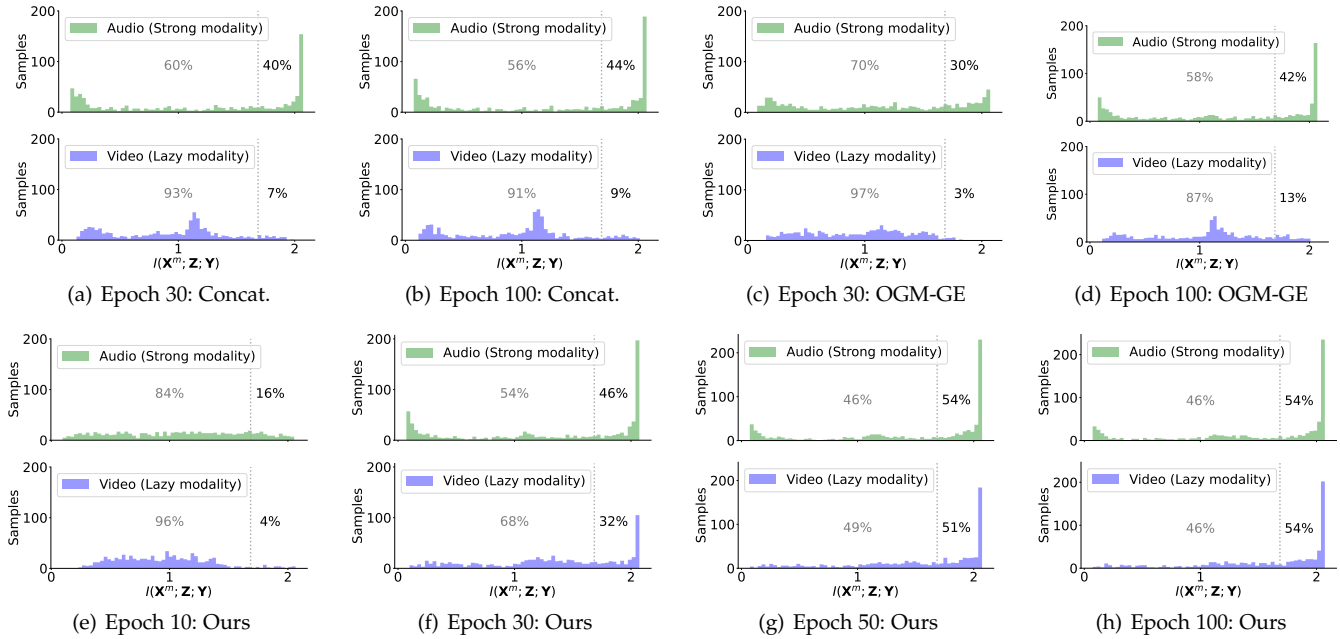
Fig. 5: The modal discrepancy ratio and recognition accuracy of the vanilla concatenation fusion method (*i.e.*, Concat), OGM-GE [2], and our method in image-text classification (*i.e.*, VT) and 2D-3D recognition (*i.e.*, 23D) on MVSA [58] and 3D MNIST datasets [59]. The red line represents the discrepancy ratio across modalities. The green, purple, and blue solid lines represent the recognition accuracy of the strong modalities, lazy modalities, and fusion in multimodal learning. The green and purple dashed lines represent the best recognition accuracy of the respective modality in unimodal learning.

critical for the effectiveness of IBML. 3) Both BIO and TCM independently contribute to more balanced optimization and improved performance relative to the vanilla concatenation baseline, confirming the effectiveness of the proposed components.

4.5 Visualization Analysis

To provide further insights into the optimization balance and performance of IBML, we conduct several visualization experiments:

1) We construct baselines by replacing the standard cross-entropy loss in OGM-GE [2] and PMR [8] with our MIB loss $\mathcal{L}_{\text{mib}}$. Figure 4 compares the modal discrepancy ratios and recognition performance among the original methods, the constructed baselines (*i.e.*, BIO+OGM-GE and BIO+PMR), and our full IBML (*i.e.*, BIO+TCM) over the training process on the CREMA-D dataset.
2) Figures 2 and 5 present comparisons of the vanilla



(a) Epoch 30: Concat. (b) Epoch 100: Concat. (c) Epoch 30: OGM-GE (d) Epoch 100: OGM-GE

(e) Epoch 10: Ours (f) Epoch 30: Ours (g) Epoch 50: Ours (h) Epoch 100: Ours

Fig. 6: The distribution comparison on the $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ of samples among different methods on CREMA-D. The green subfigures represent the strong modality, and the blue ones represent the lazy modality. Each subfigure is divided into two parts, *i.e.*, high $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ and low $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$, and the black and gray percentages represent the sample proportion of the two parts, respectively. (a) and (b) are the distribution of the vanilla concatenation fusion method (*i.e.*, Concat.) at the beginning and end of training (*i.e.*, the epoch 30 and 100), respectively. (c) and (d) are the distribution of OGM-GE [2] at epoch 30 and 100. (e) - (h) are the distribution of the complementary information of samples of our IBML at epochs 10, 30, 50, and 100.

concatenation fusion method, OGM-GE [2], MM-Pareto [12], and our IBML in terms of recognition accuracy and modal discrepancy ratio on the CREMA-D, MVSA, and 3D MNIST datasets, as illustrated in Figures 2 and 5. Notably, these datasets represent audio-visual recognition, image-text classification, and 2D-3D recognition tasks, respectively.

3) Figure 6 compares the vanilla concatenation method, OGM-GE [2], and IBML regarding the distribution on $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ of samples (approximately computed via the lower bound in Equation (11)) with different epochs on CREMA-D, thereby evaluating the performance in balanced complementary information retention of each modality.

From these visualizations, one could draw the following insights:

1) The constructed baselines achieve more balanced discrepancy ratios and improved recognition performance relative to the original methods in unimodal and joint learning. However, our IBML further outperforms these baselines, underscoring the plug-and-play capability and effectiveness of the proposed BIO for multimodal learning. Additionally, it also highlights that our TCM has a more balanced and superior optimization modulation effect compared with OGM-GE and PMR.
2) Existing methods fail to control the optimization imbalance effectively across different tasks, leading to unsatisfactory performance in lazy modalities and overall fusion. In contrast, IBML achieves a more balanced optimization between strong and lazy modalities.
3) Compared with the vanilla concatenation method, the existing heuristic balancing method (*i.e.*, OGM-GE) still struggles to balance mutual information $I(\mathbf{X}^m;\mathbf{Z};\mathbf{Y})$ between the strong and lazy modalities, which is the core component of MIB criterion. In contrast, our IBML can evenly and comprehensively excavate the mutual information of different modalities during the training process.

5 Conclusion

In this paper, we develop a Multimodal Information Balance (MIB) theory to address the imbalanced optimization problem that is widely present in multimodal joint learning. MIB reveals that its imbalance problem can be explained as the imbalance of complementary information in the modality fusion process. To this end, we define a criterion and construct a novel information-balanced Multimodal Learning (IBML) framework, which includes two novel modules (BIO and TCM) to strive for the ideal information contribution of each modality. Our method is simple, powerful, and performs well on eight widely-used multimodal datasets involving three tasks. Besides, the in-depth analysis and visualization show that our method can effectively alleviate the imbalance optimization.

References

[1] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[2] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8238–8247.

[3] Y. Wei, D. Hu, H. Du, and J.-R. Wen, "On-the-fly modulation for balanced multimodal learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] Y. Qin, Y. Sun, D. Peng, J. T. Zhou, X. Peng, and P. Hu, "Cross-modal active complementary learning with self-refining correspondence," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] Y. Feng, H. Zhu, D. Peng, X. Peng, and P. Hu, "Rono: Robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 610–11 619.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[7] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[8] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 029–20 038.

[9] X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," *arXiv preprint arXiv:2311.10707*, 2023.

[10] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 695–12 705.

[11] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *International Conference on Machine Learning*. PMLR, 2022, pp. 24 043–24 055.

[12] Y. Wei and D. Hu, "Mmpareto: boosting multimodal learning with innocent unimodal assistance," in *International Conference on Machine Learning*, 2024.

[13] Y. Wei, R. Feng, Z. Wang, and D. Hu, "Enhancing multi-modal cooperation via fine-grained modality valuation," *arXiv preprint arXiv:2309.06255*, 2023.

[14] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 214–22 224.

[15] X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 456–27 466.

[16] Z. Qin, D. Kim, and T. Gedeon, "Rethinking softmax with cross-entropy: Neural network classifier as mutual information estimator," *arXiv preprint arXiv:1911.10688*, 2019.

[17] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[18] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 247–263.

[19] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2130–2134.

[20] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[21] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 655–12 663.

[22] A. Jabri, A. Joulin, and L. Van Der Maaten, "Revisiting visual question answering baselines," in *European conference on computer vision*. Springer, 2016, pp. 727–739.

[23] T. Winterbottom, S. Xiao, A. McLean, and N. A. Moubayed, "On modality bias in the tvqa dataset," *arXiv preprint arXiv:2012.10210*, 2020.

[24] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably)," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9226–9259.

[25] H. Ni, Y. Wei, H. Liu, G. Chen, C. Peng, H. Lin, and D. Hu, "Rollingq: Reviving the cooperation dynamics in multimodal transformer," in *Forty-second International Conference on Machine Learning*.

[26] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[27] S.-H. Hwang, S. Choi, and S. E. Whang, "Midas: Misalignment-based data augmentation strategy for imbalanced multimodal learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[28] Y. Wei, S. Li, R. Feng, and D. Hu, "Diagnosing and re-learning for balanced multimodal learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–86.

[29] C. Huang, Y. Wei, Z. Yang, and D. Hu, "Adaptive unimodal regulation for balanced multimodal information acquisition," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 854–25 863.

[30] X. Gao, B. Cao, P. Zhu, N. Wang, and Q. Hu, "Asymmetric reinforcing against multi-modal representation bias," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 16, 2025, pp. 16 754–16 762.

[31] S. Wei, C. Luo, and Y. Luo, "Improving multimodal learning via imbalanced learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 2250–2259.

[32] ——, "Boosting multimodal learning via disentangled gradient learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22 879–22 888.

[33] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on neural networks*, vol. 20, no. 2, pp. 189–201, 2009.

[34] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2633–2641.

[35] J. Liu, Y. Pan, F.-X. Wu, and J. Wang, "Enhancing the feature representation of multi-modal mri data by combining multi-view information for mci classification," *Neurocomputing*, vol. 400, pp. 322–332, 2020.

[36] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 2017, pp. 263–275.

[37] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 174–11 183.

[38] C. Cui, Y. Ren, J. Pu, J. Li, X. Pu, T. Wu, Y. Shi, and L. He, "A novel approach for effective multi-view clustering with information-theoretic perspective," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[39] W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Integrating information theory and adversarial learning for cross-modal retrieval," *Pattern Recognition*, vol. 117, p. 107983, 2021.

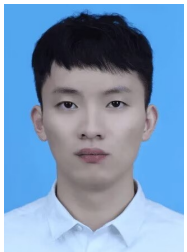
[40] Z. Wang, Y. Yang, Y. Chen, T. Yuan, M. Sermesant, H. Delingette, and O. Wu, "Mutual information guided diffusion for zero-shot cross-modality medical image translation," *IEEE Transactions on Medical Imaging*, 2024.

[41] Z. Zhang, H. Ping, P. Zhang, N. Kanakaris, X. LU, P. Bogdan, and X. Xiao, "Mihc: Multi-view interpretable hypergraph neural networks with information bottleneck for chip congestion prediction," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[42] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Transactions on Multimedia*, vol. 25, pp. 4121–4134, 2022.

[43] R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, and W. M. Wells, "Multimodal representation learning via maximization of local mutual information," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 273–283.

[44] X. Xiao, G. Liu, G. Gupta, D. Cao, S. Li, Y. Li, T. Fang, M. Cheng, and P. Bogdan, "Neuro-inspired information-theoretic hierarchical perception for multimodal learning," *arXiv preprint arXiv:2404.09403*, 2024.
[45] S. Xia, X. Zhang, H. Meng, and L. Jiao, "Ternary modality contrastive learning for hyperspectral and lidar data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.
[46] W. Su, X. Zhu, C. Tao, L. Lu, B. Li, G. Huang, Y. Qiao, X. Wang, J. Zhou, and J. Dai, "Towards all-in-one pre-training via maximizing multi-modal mutual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 888–15 899.
[47] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.
[48] M. R. Znaidi, G. Gupta, and P. Bogdan, "Secure distributed/federated learning: prediction-privacy trade-off for multi-agent system," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 97–102.
[49] R. Xu, R. Feng, S.-X. Zhang, and D. Hu, "Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
[50] D. Barber and F. Agakov, "The im algorithm: a variational approach to information maximization," *Advances in neural information processing systems*, vol. 16, no. 320, p. 201, 2004.
[51] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 875–884.
[52] X. Li, "Positive-incentive noise," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[53] X. Yu, Y. Xue, L. Zhang, L. Wang, T. Liu, and D. Zhu, "Exploring the influence of information entropy change in learning systems," *arXiv preprint arXiv:2309.10625*, 2023.
[54] K. L. Chung, "Markov chains," *Springer-Verlag, New York*, 1967.
[55] Z. Fu, Z. Mao, Y. Song, and Y. Zhang, "Learning semantic relationship among instances for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 15 159–15 168.
[56] L. Jing, E. Vahdani, J. Tan, and Y. Tian, "Cross-modal center loss for 3d cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3142–3151.
[57] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.
[58] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*. Springer, 2016, pp. 15–27.
[59] X. Xu, A. Dehghani, D. Corrigan, S. Caulfield, and D. Moloney, "Convolutional neural network for 3d object recognition using volumetric representation," in *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016, pp. 1–5.



Yang Qin received the bachelor's degree in Software Engineering from Sichuan University, Chengdu, China, in July 2021, where he is currently pursuing the Ph.D. degree with the College of Computer Science. His research interests include multimodal learning and learning with noisy correspondence.

Yanglin Feng received his bachelor's degree in Software Engineering from the College of Software Engineering at Sichuan University, Chengdu, China, in 2022. He is currently pursuing a Ph.D. in Computer Science at the College of Computer Science, Sichuan University. His research interests include multimodal learning and 3D Vision & Language.

Yuan Sun is currently pursuing the Ph.D. degree at the College of Computer, Sichuan University, Chengdu, China. He has published more than ten papers in highly regarded journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, AAAI, IJCAI, and ACM MM. His research interests include image set classification, cross-modal retrieval, and multi-view learning.

Dezhong Peng received the B.Sc. degree in applied mathematics and the M.Sc. and Ph.D. degrees in computer software and theory from the University of Electronic Science and Technology of China, Chengdu, China, in 1998, 2001, and 2006, respectively. From 2001 to 2007, he was with the University of Electronic Science and Technology of China as an Assistant Lecturer and a Lecturer. He was a Post-Doctoral Research Fellow with the School of Engineering, Deakin University, Geelong, VIC, Australia, from 2007 to 2009. He is currently a Professor with the College of Computer Science, Sichuan University, Chengdu, China. His research interests include neural networks and signal processing.

Xi Peng is currently the Cheung Kong Distinguished Professor with the College of Computer Science, Sichuan University. His current research interests include machine learning, multimedia analysis, and AI4Science. In these areas, he has co-authored around 100 articles in Nature Communications, JMLR, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, ICML, and NeurIPS.

Peng Hu received the Ph.D. degree in computer science and technology from Sichuan University, China, in 2019. He is currently an Associate Research Professor with the College of Computer Science, Sichuan University. His research interests include multi-view learning, cross-modal retrieval, and network compression. In these areas, he has authored more than 40 papers in top-tier conferences and journals.