Asymmetric Reinforcing Against Multi-Modal Representation Bias

Xiyuan Gao[1,2], Bing Cao[1,2*], Pengfei Zhu[1], Nannan Wang[2], Qinghua Hu[1]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, 300000, China
[2]The State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, 710000, China
{gaoxiyuan, caobing, zhupengfei, huqinghua}@tju.edu.cn, nnwang@xidian.edu.cn

Abstract

The strength of multimodal learning lies in its ability to integrate information from various sources, providing rich and comprehensive insights. However, in real-world scenarios, multi-modal systems often face the challenge of dynamic modality contributions, the dominance of different modalities may change with the environments, leading to suboptimal performance in multimodal learning. Current methods mainly enhance weak modalities to balance multimodal representation bias, which inevitably optimizes from a partial-modality perspective, easily leading to performance descending for dominant modalities. To address this problem, we propose an **A**symmetric **R**einforcing method against **M**ultimodal representation bias (**ARM**). Our ARM dynamically reinforces the weak modalities while maintaining the ability to represent dominant modalities through conditional mutual information. Moreover, we provide an in-depth analysis that optimizing certain modalities could cause information loss and prevent leveraging the full advantages of multimodal data. By exploring the dominance and narrowing the contribution gaps between modalities, we have significantly improved the performance of multimodal learning, making notable progress in mitigating imbalanced multimodal learning.

Introduction

Multimodal learning has emerged as a pivotal area in the field of machine learning, leveraging data from multiple sources to enhance the performance of models. This approach has been particularly transformative in applications such as image and text analysis, speech recognition, and autonomous driving, where combining visual, auditory, and textual information leads to more robust systems and makes multimodal learning an exciting frontier with significant potential (Huang et al. 2021). Despite promising yields, multimodal learning faces a critical challenge: *imbalanced learning among different modalities*. In most scenarios, partial modalities, even a single modality, may dominate the learning process, leading to insufficient learning of other modalities. Some modalities may become hard to learn due to environmental interference or limited information, leading to a multimodal bias for easier-to-learn modalities (Wu et al. 2022), and multimodal learning may degrade to unimodal learning (Huang et al. 2022).

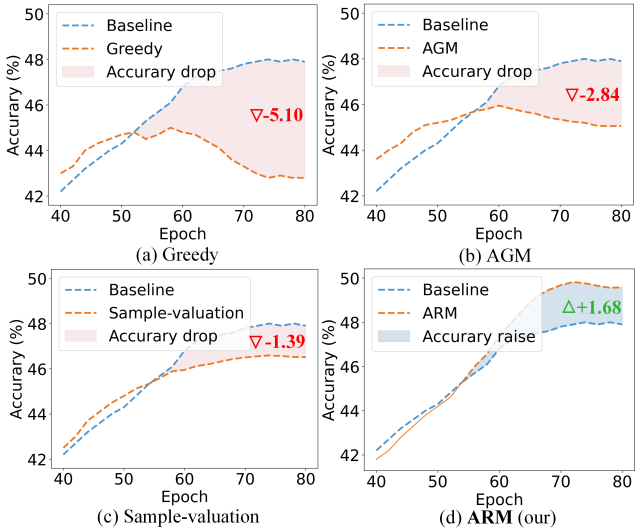


Figure 1: Accuracy curve of dominant modality compared with joint training baseline of imbalanced multimodal learning methods on Kinetics Sounds dataset. Other methods: Greedy (Wu et al. 2022), AGM(Li et al. 2023a), Sample-valuation (Wei et al. 2024).

Imbalanced learning undermines the core objective of multimodal learning, which is to harness the complementary strengths of different data formats to achieve superior performance. In recent years, many extraordinary methods have been proposed to solve this problem, including canonical correlation analysis (Sun et al. 2020), random forest (Bi et al. 2020) and ensemble learning (Livne et al. 2018). Coupled with continuously optimized large-scale datasets and algorithm innovation, deep learning methods have shown significant promise in addressing modality imbalance (Lee, Lee, and Kim 2022; Das et al. 2023). The researchers attempted to balance the multimodal learning process through methods such as gradient modulation (Peng et al. 2022; Li et al. 2023a), collaborative learning (Rahate et al. 2022), and evaluation of modality contributions (Wei et al. 2024). However, these methods alleviate the imbalance by improving the

*Corresponding author

representation of weak modalities from the uni-modal perspective alone, ignoring the connection between modalities, and not effectively utilizing all modalities. Although some methods (Zhang et al. 2024; Hu, Li, and Zhou 2022) consider cross-modal learning, they approach it from late fusion or modality preservation, without fully exploring the interrelationships between modalities, which limits their potential to improve model performance. Therefore, how to balance multimodal cooperation from a multimodal perspective remains an open question. Specifically, it is still expected to be addressed to narrow the contribution gaps between modalities and enhance the joint contribution of all modalities by exploring the interaction information between them.

To this concern, we have introduced a comprehensive valuation metric to evaluate the marginal contribution of each modality and the joint contribution of all modalities during learning for each sample. Mutual information (MI) originates from information theory used to measure the correlation between two random variables (Cover 1999). It represents the amount of information one variable contains about another and copes with capturing arbitrary dependency relationships, including linear, nonlinear, and higher-order relationships. This inspires us to use MI to measure the contribution of each modality to the learning process. To fully explore interaction information between modalities, we further utilize Conditional Mutual Information (CMI) to measure the reduction in uncertainty brought by introducing additional modalities on top of a uni-modal, thereby balancing multimodal learning without modality forgetting. Based on this, we propose an asymmetric enhancement method to dynamically alleviate imbalanced multimodal learning while maintaining the performance of dominant modalities. As shown in Fig. 1, most imbalanced multimodal learning methods exhibit dominant modality-forgetting during the training process because their optimization does not pay sufficient attention to dominant modalities, failing to maintain performance on these modalities. In contrast, based on the interrelationships between modalities, our method not only reasonably reduces the contribution disparity between modalities but also enhances the performance of each modality, overcoming the modality-forgetting. The main highlights of our study are as follows:

- We propose a mutual information-based valuation metric (MIV) to measure the marginal contribution of each modality and the joint contribution of all modalities in a sample with interrelation between modalities.
- Based on MIV, we propose an asymmetric reinforcement framework for multimodal representation bias, which dynamically narrows the contribution gaps between modalities. By continuously focusing on the dynamically changing dominance of different modalities, we mitigate modality forgetting and enhance the overall performance.
- We first reveal modality contributions from a multimodal perspective, each modality makes a positive and unique contribution to the multimodal systems. Extensive experiments validated our superiority on various multimodal classification datasets against the SOTAs.

Related Works

Imbalanced Multimodal Cooperation

Most multimodal learning often struggles with modality bias, where the dominant modality overshadows the others, leading to suboptimal performance. Recent advancements have focused on addressing this phenomenon through prototypical network (Fan et al. 2023), gradient modulation (Fu et al. 2023; Peng et al. 2022), and distilling knowledge (Pan et al. 2024; Du et al. 2021), dynamically weighing the importance of each modality based on task relevance or transferring knowledge from well-trained models, helping to mitigate imbalance. Evaluation methods (Koh et al. 2024; Yu et al. 2023), especially SHAPE (Hu, Li, and Zhou 2022) and Sample-valuation (Wei et al. 2024) novelly encourage balanced learning by improving the optimization of worse score modalities. Despite these advances, challenges remain in achieving truly balanced multimodal learning, most of these methods fall short by only enhancing weaker modalities without considering the intricate relationships between them. In contrast, we provide an asymmetric reinforcement strategy that dynamically alleviates multimodal bias based on contribution estimation without modality forgetting. This approach not only reduces the contribution disparity between modalities but also enhances overall multimodal cooperation, leading to improved performance across various multimodal classification datasets.

Mutual Information in Machine Learning

Mutual Information (MI) has been a fundamental concept in information theory and its applications in machine learning (Haghifam et al. 2020; Hadizadeh et al. 2024), which highlights the dependency between variables. In machine learning, MI is widely used for feature selection and representation learning. Early techniques (Covert et al. 2023; Stutts et al. 2023) utilized MI to identify the most relevant features for predictive modeling, improve model performance by removing redundant or irrelevant features, and allow models to focus on the most informative features. In deep learning, MI has been instrumental in unsupervised learning and generative models (Larsson et al. 2019). Techniques like InfoGAN (Chen et al. 2016) leverage MI to improve the quality of generated samples and the robustness of models. Some variational autoencoders mutations (Pan, Long, and Pan 2023) use MI to learn a latent representation that captures the underlying structure of the data while ensuring independence between latent variables. Furthermore, MI neural estimation (Kim et al. 2022) has been introduced to efficiently estimate MI between high-dimensional variables, enabling more accurate learning in complex models. Recent advancements also include using MI in knowledge distillation (Chen et al. 2023) and domain adaptation (Wen et al. 2024), where understanding the information flow between different domains or causal variables is crucial. Overall, mutual information continues to be a powerful tool in enhancing the capabilities of machine learning models, driving us to use mutual information to measure modal benefits. To the best of our knowledge, we for the first time utilize mutual information to handle imbalanced multimodal learning.

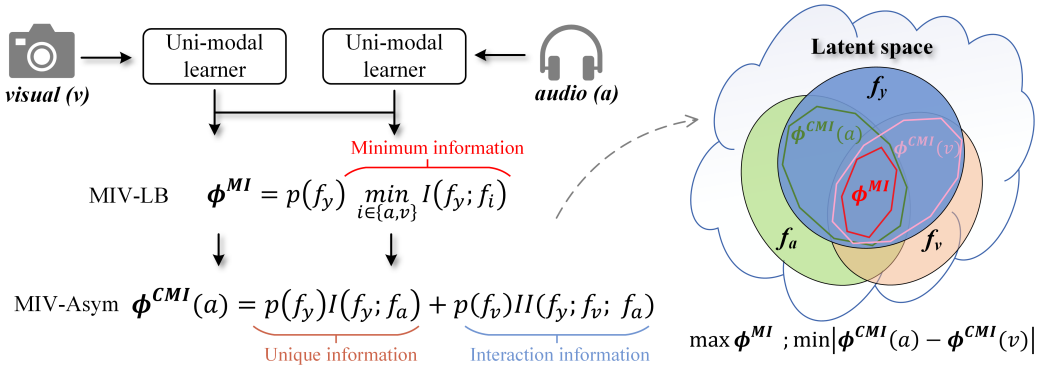


Figure 2: **Left:** The Lower Bound joint contribution (MIV-LB) of all modalities and the Asymmetric marginal contribution (MIV-Asym) of each modality are estimated by ϕ^{MI} and ϕ^{CMI}, respectively, serving as the basis for asymmetric reinforcement. $f_{\mathcal{Y}}$ is feature-level fusion result, p is the accurate production. **Right:** Representation of features in the latent space. We minimize the diversities in ϕ^{CMI} to balance multimodal learning while maximizing ϕ^{MI} to enhance multimodal performance.

Methods

Preliminary

In an interactive system, we can obtain partial information about one variable *X* by observing another variable *Y*, thereby reducing the uncertainty of the former. The extent of this uncertainty reduction can be considered a measure of contribution and can be quantified using Mutual Information (MI). Using the basic relationship between the MI and entropy $H(\cdot)$ (Cover 1999), the algorithm for MI can be defined as the individual entropy of *X*, minus the conditional entropy of *X* given *Y*. Following this approach, we can derive the formula for MI and Normalized MI (NMI) as follows:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} \mathcal{P}(x,y)\log\frac{\mathcal{P}(x,y)}{\mathcal{P}(x)\mathcal{P}(y)}, \quad (1)$$

$$NMI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}. \quad (2)$$

Considering a multimodal classification task, a sample with *m* modalities is represented as $\mathcal{X} = \{x^1, x^2, \ldots, x^m\}$, which can be regarded as a multimodal pair, and y is the ground truth label of sample $\mathcal{X}$. Denote a uni-modal encoder as $\mathcal{E}(\cdot)$, the classification head as $\mathcal{H}(\cdot)$. The feature of the *i*-th modality extracted by the encoder is $f_{x^i} = \mathcal{E}(x^i)$. When taking $\mathcal{X}$ as the input for multimodal learning, the feature-level fusion output is $f_{\mathcal{Y}} = \cup f_{x^i}$, $x^i \in \mathcal{X}$, the final prediction is $\hat{y} = \mathcal{H}(f_{\mathcal{Y}})$. Notably, in this multimodal classification task, the features f_{x^i} of the multimodal pair $\mathcal{X}$ are fused to obtain $f_{\mathcal{Y}}$. Subsequently, $f_{\mathcal{Y}}$ is used to make the final prediction $\hat{y}$, and the parameters of $\mathcal{E}(\cdot)$ are optimized by backpropagation based on $\hat{y}$, that is, $f_{\mathcal{Y}}$ further applied to f_{x^i}. Hence, a system characterized by the mutual interaction between f_{x^i} and $f_{\mathcal{Y}}$ is constituted.

Valuation Metric without Modality Forget

When the number of variables in an interactive system increases, such as in multimodal learning, where features $f_{\mathcal{X}} = \{f_{x^1}, f_{x^2}, \ldots, f_{x^m}\}$ from *m* modalities jointly influence the fusion result $f_{\mathcal{Y}}$, using mutual information can become challenging. Inspired by exhaustively decomposing in a multivariate system (Williams and Beer 2010), even in cases where multiple source variables jointly influence a single variable, we can still compute the MI: $I(f_{\mathcal{Y}}; f_{x^i})$ for each $f_{x^i} \in f_{\mathcal{X}}$ with $f_{\mathcal{Y}}$ separately. Notably, Eq. (1) is non-negative, so it has a positive contribution to learning each modality. The MI between $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$ can be expressed as:

$$I(f_{\mathcal{Y}}; f_{\mathcal{X}}) = \sum_{\hat{y}\in f_{\mathcal{Y}}}\sum_{x\in f_{\mathcal{X}}} \mathcal{P}(x,\hat{y})\log\frac{\mathcal{P}(x,\hat{y})}{\mathcal{P}(x)\mathcal{P}(\hat{y})}, \quad (3)$$

through observing $f_{\mathcal{Y}}$, the distribution of $f_{\mathcal{X}}$ changes from $\mathcal{P}(x)$ to $\mathcal{P}(x|\hat{y})$, we can capture the knowledge of $f_{\mathcal{X}}$ after the observation, the positive contribution in Eq. (3) is where predicting the ground truth label y, that is:

$$I(f_{\mathcal{Y}} = y; f_{\mathcal{X}}) = \sum_{\boldsymbol{x}\in f_{\mathcal{X}}} \mathcal{P}(\boldsymbol{x}|y)\log\frac{\mathcal{P}(y \mid \boldsymbol{x})}{\mathcal{P}(y)}. \quad (4)$$

Theorem 1. *In multimodal learning with m modalities, each modality can provide a* ***positive*** *and* ***unique*** *contribution to accurate prediction. i.e.,* $I(f_{\mathcal{Y}} = y; f_{x^i}) \neq I(f_{\mathcal{Y}} = y; f_{x^j})$, *for any* $x^i, x^j \in \mathcal{X}, i \neq j$. *Naturally, neglecting the learning of any modality will result in information loss.* (The specific theoretical proof process is provided in the Appendix.)

Based **Theorem 1**, we propose a valuation metric to measure the marginal contribution of each modality in a sample $\mathcal{X}$, i.e. $\phi(x^i)$ and further derive the joint contribution of all modalities in that sample, i.e. $\phi(\mathcal{X})$. This serves as the foundation for asymmetric enhancement.

Lower bound of joint contribution $\phi(\mathcal{X})$. NMI between uni-modal and fused feature $NMI(f_{\mathcal{Y}}; f_{x^i})$ can be understood as the expected contribution value of all possible predictions from $f_{\mathcal{Y}}$ when f_{x^i} is given, and it can be expressed as Eq. (5). For clarity, we use I to represent NMI.

$$I\left(f_{\mathcal{Y}}; f_{x^i}\right) = \sum_{\hat{y}}^{\mathrm{N}} p\left(f_{\mathcal{Y}} \to \hat{y}\right) I\left(f_{\mathcal{Y}}; f_{x^i}\right), \quad (5)$$

where $p(f_{\mathcal{Y}} \to \hat{y})$ represent the probability that $f_{\mathcal{Y}}$ makes the final prediction of class $\hat{y}$, N is the number of categories. As we adopt *Softmax*, $\sum_{\hat{y}}^{\mathrm{N}} p\left(f_{\mathcal{Y}} \to \hat{y}\right) = 1$. Therefore, based on the MI, the contribution of the model's accurate prediction provided by i-th modality can be written as:

$$\phi^{MI}\left(x^i\right) = p\left(f_{\mathcal{Y}} \to y\right) I\left(f_{\mathcal{Y}}; f_{x^i}\right). \tag{6}$$

Similarly, observing j-th modality ($j \neq i$) can also contribute to an extent that $f_{\mathcal{Y}}$ makes the accurate prediction y. Hence, the lower bound of joint contribution for all modalities in sample $\mathcal{X}$ is:

$$\phi^{MI}(\mathcal{X}) = p\left(f_{\mathcal{Y}} \to y\right) \min_{i \in \{1,\ldots m\}} I\left(f_{\mathcal{Y}}; f_{x^i}\right). \tag{7}$$

It represents the minimum contribution value that each modality can provide for the model's accurate prediction. ϕ^{MI} has several properties: Firstly, its value range is [0,1]. Secondly, ϕ^{MI} is less than or equal to $I(f_{\mathcal{Y}}; f_{x^i})$ for all $i \leq m$. Finally, in the training phase, by incorporating ϕ^{MI} into the loss function and using gradient descent to maximize ϕ^{MI}, thus each iteration moves towards increasing mutual information, ensuring the convergence of the lower bound.

Estimating marginal contribution $\phi(x^i)$. Although we defined the lower bound joint contribution of sample $\mathcal{X}$, the interrelationships between modalities are ignored, which prevents us from fully leveraging the advantages of multimodal learning. As one would hope, given the presence of variable Z, the impact of introducing an additional variable Y on X can be measured using Conditional Mutual Information (CMI). The formulas for CMI and Normalized CMI (NCMI) are as follows:

$$\begin{aligned} CMI_{(X;Y|Z)} &= \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} \mathcal{P}(x, y, z) \log \frac{\mathcal{P}(x, y|z)}{\mathcal{P}(x|z)\mathcal{P}(y|z)}, \\ &= \mathbb{E}_Z D_{KL}\left[\mathcal{P}(x, y|z) \| \mathcal{P}(x|z)\mathcal{P}(y|z)\right] \end{aligned} \tag{8}$$

$$NCMI_{(X;Y|Z)} = \frac{CMI(X;Y|Z)}{\sqrt{H(X|Z)H(Y|Z)}}. \tag{9}$$

In a complete modality set $\mathcal{X}$, when we choose the x^i and x^j to calculate MI with the fusion result separately, the mutual information of f_{x^i} is $I(f_{\mathcal{Y}}; f_{x^i})$, and the conditional mutual information of f_{x^j} given f_{x^i} is $I(f_{\mathcal{Y}}; f_{x^j}|f_{x^i})$, vice versa. Consequently, the change in contribution value that modality x^j causes to modality x^i is the Interaction Information (II):

$$II(f_{\mathcal{Y}}; f_{x^j}; f_{x^i}) = I(f_{\mathcal{Y}}; f_{x^j}) - NCMI(f_{\mathcal{Y}}; f_{x^j}|f_{x^i}). \tag{10}$$

With Eq. (10), we can estimate the marginal contribution of x^i based on considering all modalities as follows:

$$\begin{aligned} \phi^{CMI}(x^i) = &\, p\left(f_{\mathcal{Y}} \to y\right) I\left(f_{\mathcal{Y}}; f_{x^i}\right) \\ &+ \sum_{j \neq i}^{m} p(f_{x^j} \to y) II(f_{\mathcal{Y}}; f_{x^j}; f_{x^i}), \end{aligned} \tag{11}$$

where $p(f_{x^j} \to y)$ can be regarded as a dynamic modality-specific weight of j-th modality, which can heighten the model's robustness (Yang et al. 2024). Furthermore, the joint contribution of the complete modality set from sample $\mathcal{X}$ can be expressed as:

$$\phi^{CMI}(\mathcal{X}) = \frac{1}{m} \sum_{i=1}^{m} \phi^{CMI}(x^i). \tag{12}$$

ϕ^{CMI} has several advantages: Firstly, it considers the impact of each modality from sample $\mathcal{X}$, ensuring that there is no modality omission during learning. Secondly, its value range is [0, m], allowing it to be flexibly incorporated into loss functions or regularization as an optimization technique. Finally, averaging reasonably reflects the salient characteristics of the overall modalities, preventing the landslide victory of certain modalities while suppressing the occurrence of outliers.

Asymmetric Reinforcement Strategies

Dynamic Feature-level Fusion. Considering real-world factors, due to the primacy effect, the effect of the first term in Eq. (11) will be amplified. In other words, $\phi^{CMI}(x^i)$ reflects the importance to accurate prediction of i-th modality. We can use this as the specific-modal fusion weight during the fusion phase. Generally, higher $\phi^{CMI}(x^i)$ values represent more positive impacts on the model, thus the Fusion Weight (FW) of i-th modality can be denoted as:

$$FW^i = \frac{\phi^{CMI}(x^i)}{\phi^{CMI}(\mathcal{X})}, \tag{13}$$

where FW^i works during the training phase and will take effect in the next epoch.

Balanced Min-Max Loss. Examining the expression of Eq. (7), (12), it is evident that ϕ^{MI} and ϕ^{CMI} represent the minimum contribution and comprehensive contribution that complete modalities for the model's accurate prediction, respectively. For the former, maximizing ϕ^{MI} enables the model to learn the most beneficial aspect of each modality for accurate prediction, and for the latter, we can use the Mean Absolute Error (MAE) to minimize $MAE(\phi^{CMI})$, thereby narrowing the marginal contribution gap between modalities.

$$\mathcal{L}_{\phi^{MI}} = 1 - \phi^{MI}(\mathcal{X}), \tag{14}$$

$$\mathcal{L}_{\phi^{CMI}} = \frac{\sum_{i=1}^{m} \left|\phi^{CMI}(x^i) - \phi^{CMI}(\mathcal{X})\right|}{\phi^{CMI}(\mathcal{X})}. \tag{15}$$

It should be noted that Eq. (14) cannot directly participate in the gradient backward process of gradient descent optimization since the min function is not globally differentiable. To this end, we use smooth approximation (Nielsen and Sun 2016) to make it differentiable:

$$\min_{i \in \{1,\ldots m\}} I^i = \max_{i \in \{1,\ldots m\}} (-I^i) \approx \log(\sum_{i=1}^{m} e^{-I^i}). \tag{16}$$

The overall loss function of ARM is formulated as Eq. (17), where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, λ_1 and λ_2 are trade-off parameters.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{\phi^{MI}} + \lambda_2 \mathcal{L}_{\phi^{CMI}}. \tag{17}$$

Dynamic Sample-level Re-sample. Following the analysis in **Theorem 1** and specific theoretical in (Wei et al. 2024), enhancing the discriminative ability of lower-contribution modality can expand its contribution. We propose to resample all modalities of lower joint contribution sample $\mathcal{X}$ more frequently during training. After each modality valuation by MIV, we can dynamically determine the re-sampling frequency with ϕ^{CMI} to enhance contribution, where re-sample frequency of sample $\mathcal{X}$ is:

$$s(\mathcal{X}) = \mathcal{F}_s(\phi^{CMI}(\mathcal{X})), \tag{18}$$

where $\mathcal{F}_s$ is a monotonically decreasing function, the lower-contribution sample $\mathcal{X}$ is re-trained with a resample frequency inversely proportional to its joint contribution. It is worth noting that different from (Wei et al. 2024), our re-sampling strategy is from a multimodal perspective, which dynamically adjusts the sampling frequency of all modalities in $\mathcal{X}$. This ensures that no information is lost during training, while the loss function $\mathcal{L}_{\phi^{MI}}$ guarantees targeted learning for lower-contribution modalities.

Experiments

Datasets and Implementation Details

Kinetic Sounds (KS) (Arandjelovic and Zisserman 2017) is a specifically designed action recognition dataset for research in audio-visual learning, particularly focusing on the relationship between actions and corresponding sounds. KS is composed of YouTube videos; all videos are cropped to within 10 seconds around the action. KS includes approximately 23k video clips with 31 categories.

UCF-51 is a subset of UCF-101 (Soomro, Zamir, and Shah 2012) with two modalities, RGB and optical flow, containing 6,845 video clips across 51 diverse action categories. Mostly sourced from YouTube, it features varying conditions such as different camera angles and lighting, making it challenging and realistic for real-world applications.

UPMC Food-101 (Wang et al. 2015) is a comprehensive dataset for food recognition, consisting of 101,000 images accompanied by corresponding texts across 101 food categories. Each category includes 750 images for training and 250 images for testing.

Implementation Details. Unless otherwise specified, ResNet-18 is used as the backbone in the experiments and trained from scratch. Encoders used for UCF-51 are ImageNet pre-trained. For Food-101, a ViT-based model is used as the vision encoder, and a BERT-based model is used as the text encoder by the pre-trained. Before modality valuation, a warm-up stage is employed for all experiments. During training, we use Stochastic Gradient Descent (SGD) with a batch size of 64. We set the initial learning rate, weight decay, and momentum parameters to 10^{-3}, 5×10^{-4}, and 0.9, respectively. The experiments are conducted on Huawei Atlas 800 Training Server with CANN and NVIDIA 4090 GPU. More details of implementation and experiment analysis are provided in the Appendix.

Model	KS (Acc.)	UCF-51 (Acc.)
Concatenation	59.61	68.23
Summation	59.53	67.62
OGM-GE (CVPR 2022)	60.70	71.66
Greedy (ICML 2022)	59.86	71.53
QMF (ICML 2023)	63.78	73.48
PMR (CVPR 2023)	63.86	74.80
Sample-val. (CVPR 2024)	<u>65.33</u>	75.12
Modality-val. (CVPR 2024)	65.10	74.39
MLA (CVPR 2024)	65.21	**76.01**
ARM	**66.52**	<u>75.60</u>

Table 1: Accuracy of imbalanced multimodal learning methods, where bold and underline represent the best and runner-up respectively. Results are reported in percentage (%).

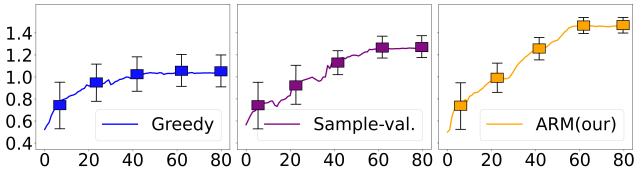


Figure 3: Comparison of the narrowing trend of uni-modality contribution gaps on UCF-51. The horizontal axis is *Epoch*, and the vertical axis is *joint contribution*.

Comparison with Imbalanced Multimodal Learning Methods

In this section, we compared ARM with advanced imbalanced multimodal learning methods to answer **Q1:** *How does ARM narrow the modality contribution gap?*

Fig. 3 illustrates the trend of narrowing contribution gaps across different methods. Traditional methods, like Greedy, have shown limited improvement in closing the contribution disparity between modalities, with only a slight narrowing in the contribution gap as training progresses. Sample-valuation demonstrates more consistent shrink, yet the gap remains noticeable across epochs. In contrast, ARM achieves a marked and consistent reduction in modality contribution gaps, indicating a more balanced learning process. This consistent improvement shows ARM's ability to maintain equitable contribution from all modalities, which is crucial for robust multimodal learning.

Table 1 further reinforces this conclusion. ARM consistently outperforms other state-of-the-art methods, i.e., Greedy (Wu et al. 2022), OGM-GE (Peng et al. 2022), QMF (Zhang et al. 2023), PMR (Fan et al. 2023), Sample-valuation, Modality-valuation (Wei et al. 2024), and MLA (Zhang et al. 2024), achieving the competitive accuracy scores of 66.52% and 75.60%, respectively. Other approaches, like QMF and PMR, show decent performance but still fall short in balancing modality contributions, leading to suboptimal accuracy. Due to the different design focus, MLA performs better in handling temporal optical flow data in the UCF-51 dataset. Sample-valuation achieves competitive results but cannot match the balance achieved by ARM, which is evident from the joint contribution trends shown

Model	KS (Acc.)	Food-101 (Acc.)
Concatenation	59.61	82.38
Summation	59.53	82.63
Decision fusion	60.12	83.71
FiLM (AAAI 2018)	59.33	82.34
BiGated (AAAI 2018)	60.79	86.71
Dynamic Fusion (CVPR 2023)	63.21	90.83
PMF (ICCV 2023)	64.33	91.56
TransFusion (ICLR 2024)	65.40	91.22
ARM	**66.52**	**93.36**

Table 2: Comparison with multimodal fusion methods.

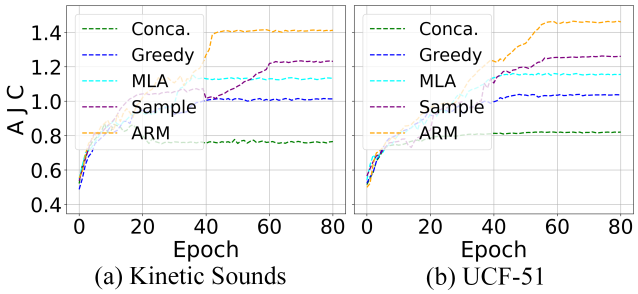
in Fig. 3. The advantage of ARM lies in its dual focus: minimizing modality imbalances while maximizing overall performance. By effectively narrowing the contribution gaps between modalities, ARM prevents any single modality from dominating or being neglected, leading to a more cohesive and effective multimodal representation.

Comparison with Multimodal Fusion Methods

Table 2 compares the performance of various multimodal fusion methods on two datasets to answer **Q2:** *Can the proposed modules (e.g., dynamic feature-level fusion) effectively improve performance?*

Concatenation and Summation are baseline methods that simply merge the feature vectors, yielding moderate performance. More advanced techniques such as FiLM (Perez et al. 2018) and BiGated (Kiela et al. 2018) introduce interaction between modalities through modulation or gating mechanisms, resulting in eligible accuracy compared with the baseline. Dynamic Fusion (Xue and Marculescu 2023) incorporates adaptive fusion strategies, which inspire our dynamic feature-level fusion, adjusting how the modalities are combined during inference, which leads to substantial improvements, especially on the Food-101 dataset.

Among the recently proposed methods, PMF (Li et al. 2023b) and TransFusion (Imfeld et al. 2023) demonstrate the power of more sophisticated fusion techniques. PMF achieves strong performance by effectively managing modality-specific features. TransFusion, a transformer-based model, further refines this by better capturing the complex interactions between modalities, achieving runner-up results. Our ARM outperforms all other models on both datasets, achieving 66.52% accuracy on KS and 93.36% on Food-101, which is a significant improvement, particularly evident on the KS dataset, where it exceeds the runner-up by over 1 percentage point. ARM's success is attributed to its advanced asymmetric reinforcement strategy, which dynamically balances the learning from each modality, preventing the model from being biased toward the dominant modality. This ensures that both audio and visual information are utilized effectively, leading to superior performance in challenging multimodal tasks. Compared with the competing methods, ARM's ability to maintain high accuracy across different datasets demonstrates its robustness and adaptability, making it a standout choice for multimodal fusion tasks.



(a) Kinetic Sounds (b) UCF-51

Figure 4: Average joint contribution (AJC) of all modalities overall training samples during training for Greedy, MLA, Sample-valuation, and our ARM on KS and UCF-51.

Analysis of Modality Forget & Multimodal Cooperation

We report the results of a single modality and a combination of all modalities and further display the improvement of multimodal cooperation to answer **Q3:** *Compared to prior multimodal learning approaches, can ARM overcome modality forget and optimize multimodal cooperation?*

Modality Forget. Table 3 compares the performance of various models across multiple datasets, highlighting results for different modalities and their multimodal cooperation. Several models in the comparison exhibit a notable modality forget phenomenon, where optimizing for one weaker modality leads to a decrease in the performance of the dominant modality and achieves suboptimal results in the overall multimodal performance. For instance, on the KS dataset, models like BiGated and PMF show significant drops in performance for the visual modality compared to the audio, which in turn negatively impacts their multimodal accuracy. This trend is also observed on UCF-51, where models fail to balance the learning of RGB and optical flow modalities, leading to lower overall performance. The Sample-valuation model also shows a considerable drop across both visual and textual modalities on Food-101, which further highlights the issue of modality forgetting. Our proposed ARM consistently outperforms other models across all datasets, achieving the highest accuracy in both single and multimodal scenarios. Notably, ARM excels in preventing modality forgetting, as demonstrated by its superior performance across different modalities and their combinations.

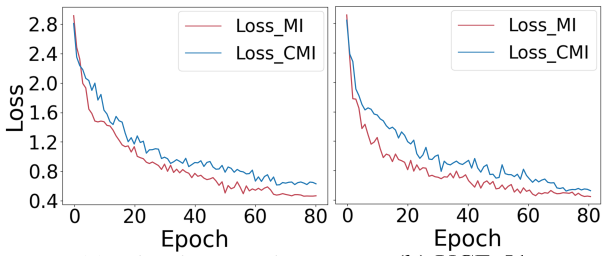
Multimodal Cooperation. Fig. 4 illustrates the progression of multimodal joint contributions over epochs compared with different models. The performance of the other methods indicates a relatively slower and less stable increase in the multimodal joint contribution over time. Greedy demonstrates some improvement but plateaus early, indicating that it struggles to maintain steady enhancement of multimodal cooperation. Sample and MLA show better performance than Concatenation and Greedy but still fall short compared to ARM, as they are unable to fully exploit the joint potential of multimodal learning. ARM exhibits a consistent and substantial increase in the multimodal average

Dataset		Conact.	Sum	BiGated	PMF	QMF	Sample	MLA	**ARM**
KS	(⋆) Audio	47.35	46.21	44.11 (↓)	45.82 (↓)	47.56	46.02 (↓)	49.20	**49.95**
	Video	23.65	22.78	22.08 (↓)	25.65	36.82	42.67	41.30	**44.86**
	Mutli	59.61	59.53	60.79	64.33	63.78	65.33	65.21	**66.52**
UCF-51	(⋆) RGB	60.13	59.80	57.39 (↓)	58.13 (↓)	56.20 (↓)	57.01 (↓)	**64.81**	63.29
	OF	29.62	28.81	25.67 (↓)	36.21	40.51	42.33	41.26	**43.19**
	Mutli	68.23	67.62	70.87	72.09	73.48	75.12	**76.01**	75.60
Food-101	Image	30.85	31.66	48.87	59.21	66.39	**73.49**	71.58	72.36
	(⋆) Text	81.68	80.84	78.51 (↓)	79.66 (↓)	82.10	84.43	86.42	**86.86**
	Mutli	82.38	82.63	86.71	91.56	91.67	90.85	93.31	**93.36**

Table 3: Comparison results on audio-video, RGB-optical flow, and image-text datasets. The performance of a single modality and the results of combining all modalities ("multiple") are listed. ⋆ denotes the dominant modality and ↓ indicates a performance drop compared with Concatenation or Sum baseline.

$\mathcal{L}_{CE}$	$\mathcal{L}_{\phi^{MI}}$	$\mathcal{L}_{\phi^{CMI}}$	KS	UCF-51	Food-101
✓			63.88	70.03	88.52
✓	✓		64.20	73.56	91.25
✓		✓	65.19	72.10	89.78
✓	✓	✓	**66.52**	**75.60**	**93.36**

Table 4: Ablation study of loss function.



(a) Kinetic Sounds (b) UCF-51

Figure 5: Curve of Balanced Min-Max Loss: the values are obtained from 5 training processes with the same initiations.

contribution, the chart shows that ARM not only achieves a higher overall contribution but also demonstrates a stable and continuous growth trend, indicating its robustness in integrating information from various modalities.

ARM's success can be attributed to its dynamic asymmetric reinforcement, which effectively balances the learning from each modality based on their importance. By dynamically adjusting the contribution of each modality and interaction with others, ARM ensures that no single modality dominates at the expense of others and allows ARM to maximize the joint contribution of all modalities, leading to superior performance in multimodal learning.

The Effectiveness of Loss Function

This section answers the question: **Q4:** *Does our proposed Balanced Min-Max loss progress as expected?*

Fig. 5 demonstrates the effectiveness of the proposed $\mathcal{L}_{\phi^{MI}}$ and $\mathcal{L}_{\phi^{CMI}}$ in improving overall multimodal performance and alleviating imbalanced learning between modalities, respectively. The loss curves for both the KS and UCF-51 datasets show that incorporating the Balanced Min-Max loss consistently improves the overall model performance by ensuring better multimodal cooperation, leading to faster convergence and lower loss values. Table 4 further validates these observations with an ablation study. When only the $\mathcal{L}_{\phi^{MI}}$ is added, there is a noticeable increase in accuracy compared to the baseline (row 1). Additionally, the inclusion of the $\mathcal{L}_{\phi^{CMI}}$ specifically addresses modality imbalance by reducing the learning disparity between modalities. This is particularly important in scenarios where dominant modalities may overshadow weaker ones. The combined use of both $\mathcal{L}_{\phi^{MI}}$ and $\mathcal{L}_{\phi^{CMI}}$ achieves the best performance, demonstrating that our approach not only enhances overall accuracy but also maintains balanced contributions across all modalities. This synergy between the two losses highlights the strength of our method in multimodal learning.

Conclusion

In this paper, we introduce a valuation metric to evaluate the marginal and joint contribution of modalities in a sample with a theoretical analysis of mutual information. Based on this, an asymmetric enhancement method named ARM is proposed to improve imbalanced multimodal learning while preventing modality forgetting. This provides a potential approach for balancing multimodal learning in real-world applications. Besides, there are some further discussions.

Universality of Mutual Information. Our method calculates mutual information after feature reduction, and the data dimension is lower, so we can directly calculate the marginal distribution and joint distribution. But when processing continuous data, discretization or kernel density estimation methods are required. These methods are relatively complex to implement and may lead to different results.

Natural Conflict in Multimodal. Multimodal data may contain some inherent conflicts. For example, for an RGB-Infrared sample *person* in foggy environments, two modalities may make vastly different predictions. Although ARM copes with mitigating modality conflicts by reducing the impact of modalities with incorrect predictions, it does not fundamentally resolve such conflicts. Therefore, it is expected to consider this natural conflict in the future work.

Acknowledgments

This work was sponsored by National Science and Technology Major Project (No. 2022ZD0116500), National Natural Science Foundation of China (No.s 62476198, 62436002, U23B2049, 62222608, 62106171, and 61925602), Tianjin Natural Science Funds for Distinguished Young Scholar (No. 23JCJQJC00270), the Zhejiang Provincial Natural Science Foundation of China (No. LD24F020004), and CCF-Baidu Open Fund. This work was also sponsored by CAAI-CANN Open Fund, developed on OpenI Community.

References

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, 609–617.

Bi, X.-a.; Hu, X.; Wu, H.; and Wang, Y. 2020. Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. *IEEE Journal of Biomedical and Health Informatics*, 24: 2973–2983.

Chen, M.; Xing, L.; Wang, Y.; and Zhang, Y. 2023. Enhanced multimodal representation learning with cross-modal kd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11766–11775.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.

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

Covert, I. C.; Qiu, W.; Lu, M.; Kim, N. Y.; White, N. J.; and Lee, S.-I. 2023. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, 6424–6447. PMLR.

Das, A.; Das, S.; Sistu, G.; Horgan, J.; Bhattacharya, U.; Jones, E.; Glavin, M.; and Eising, C. 2023. Revisiting modality imbalance in multimodal pedestrian detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 1755–1759. IEEE.

Du, C.; Li, T.; Liu, Y.; Wen, Z.; Hua, T.; Wang, Y.; and Zhao, H. 2021. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*.

Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20029–20038.

Fu, J.; Gao, J.; Bao, B.-K.; and Xu, C. 2023. Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hadizadeh, H.; Yeganli, S. F.; Rashidi, B.; and Bajić, I. V. 2024. Mutual Information Analysis in Multimodal Learning Systems. *arXiv preprint arXiv:2405.12456*.

Haghifam, M.; Negrea, J.; Khisti, A.; Roy, D. M.; and Dziugaite, G. K. 2020. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33: 9925–9935.

Hu, P.; Li, X.; and Zhou, Y. 2022. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Huang, Y.; Lin, J.; Zhou, C.; Yang, H.; and Huang, L. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, 9226–9259. PMLR.

Imfeld, M.; Graldi, J.; Giordano, M.; Hofmann, T.; Anagnostidis, S.; and Singh, S. P. 2023. Transformer fusion with optimal transport. *arXiv preprint arXiv:2310.05719*.

Kiela, D.; Grave, E.; Joulin, A.; and Mikolov, T. 2018. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kim, J.-H.; Kim, Y.; Lee, J.; Yoo, K. M.; and Lee, S.-W. 2022. Mutual information divergence: A unified metric for multimodal generative models. *Advances in Neural Information Processing Systems*, 35: 35072–35086.

Koh, J. Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M. C.; Huang, P.-Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; and Fried, D. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

Larsson, M.; Stenborg, E.; Toft, C.; Hammarstrand, L.; Sattler, T.; and Kahl, F. 2019. Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 31–41.

Lee, H. K.; Lee, J.; and Kim, S. B. 2022. Boundary-focused generative adversarial networks for imbalanced and multi-modal time series. *IEEE Transactions on Knowledge and Data Engineering*, 34: 4102–4118.

Li, H.; Li, X.; Hu, P.; Lei, Y.; Li, C.; and Zhou, Y. 2023a. Boosting Multi-modal Model Performance with Adaptive Gradient Modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22214–22224.

Li, Y.; Quan, R.; Zhu, L.; and Yang, Y. 2023b. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2604–2613.

Livne, M.; Boldsen, J. K.; Mikkelsen, I. K.; Fiebach, J. B.; Sobesky, J.; and Mouridsen, K. 2018. Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke. *Stroke*, 49: 912–918.

Nielsen, F.; and Sun, K. 2016. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18: 442.

Pan, W.; Long, F.; and Pan, J. 2023. ScInfoVAE: interpretable dimensional reduction of single cell transcription data with variational autoencoders and extended mutual information regularization. *BioData Mining*, 16: 17.

Pan, Y.; Jiang, J.; Jiang, K.; and Liu, X. 2024. Disentangled-Multimodal Privileged Knowledge Distillation for Depression Recognition with Incomplete Multimodal Data. In *ACM Multimedia*.

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

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rahate, A.; Walambe, R.; Ramanna, S.; and Kotecha, K. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81: 203–239.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Stutts, A. C.; Erricolo, D.; Ravi, S.; Tulabandhula, T.; and Trivedi, A. R. 2023. Mutual information-calibrated conformal feature fusion for uncertainty-aware multimodal 3d object detection at the edge. *arXiv preprint arXiv:2309.09593*.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8992–8999.

Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Wei, Y.; Feng, R.; Wang, Z.; and Hu, D. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27338–27347.

Wen, L.; Chen, S.; Xie, M.; Liu, C.; and Zheng, L. 2024. Training multi-source domain adaptation network by mutual information estimation and minimization. *Neural Networks*, 171: 353–361.

Williams, P. L.; and Beer, R. D. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Wu, N.; Jastrzebski, S.; Cho, K.; and Geras, K. J. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, 24043–24055. PMLR.

Xue, Z.; and Marculescu, R. 2023. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2575–2584.

Yang, Z.; Wei, Y.; Liang, C.; and Hu, D. 2024. Quantifying and Enhancing Multi-modal Robustness with Modality Preference. In *The Twelfth International Conference on Learning Representations*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J. T.; and Peng, X. 2023. Provable Dynamic Fusion for Low-Quality Multimodal Data. In *International Conference on Machine Learning*.

Zhang, X.; Yoon, J.; Bansal, M.; and Yao, H. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27456–27466.