Enhancing Multimodal Cooperation via Sample-level Modality Valuation

Yake Wei[1], **Ruoxuan Feng**[1], **Zihe Wang**[1,2], **Di Hu**[1,2,*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
{yakewei, fengruoxuan, wang.zihe, dihu}@ruc.edu.cn

Abstract

One primary topic of multimodal learning is to jointly incorporate heterogeneous information from different modalities. However, most models often suffer from unsatisfactory multimodal cooperation, which cannot jointly utilize all modalities well. Some methods are proposed to identify and enhance the worse learnt modality, but they are often hard to provide the fine-grained observation of multimodal cooperation at sample-level with theoretical support. Hence, it is essential to reasonably observe and improve the fine-grained cooperation between modalities, especially when facing realistic scenarios where the modality discrepancy could vary across different samples. To this end, we introduce a sample-level modality valuation metric to evaluate the contribution of each modality for each sample. Via modality valuation, we observe that modality discrepancy indeed could be different at sample-level, beyond the global contribution discrepancy at dataset-level. We further analyze this issue and improve cooperation between modalities at sample-level by enhancing the discriminative ability of low-contributing modalities in a targeted manner. Overall, our methods reasonably observe the fine-grained uni-modal contribution and achieve considerable improvement. The source code and dataset are available at https://github.com/GeWu-Lab/Valuate-and-Enhance-Multimodal-Cooperation.

1. Introduction

Humans are surrounded by messages of multiple senses, including vision, auditory and tactile, bringing us a comprehensive perception. Inspired by this multi-sensory integration phenomenon, learning from multimodal data has raised attention in recent years. Recent researchers have well summarized the wide range of applications of multimodal learning and looked at its future development [32]. One primary topic in multimodal learning is how to jointly incorporate multiple heterogeneous information. In the early, researchers tried to achieve the union of multiple modalities via different perspectives, like probabilistic theory based dynamic Bayesian networks [18] and multimodal Restricted Boltzmann Machines inspired by thermodynamic [19]

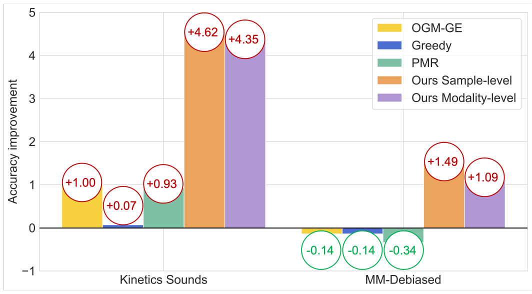
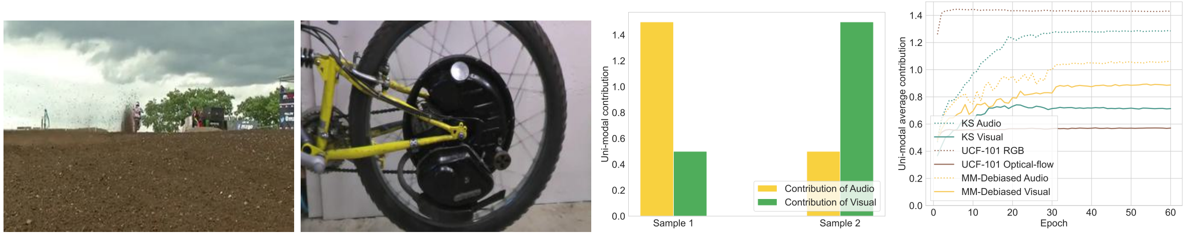


Figure 1. Accuracy improvement compared with joint training baseline of imbalanced multimodal learning methods, on Kinetics Sounds and our proposed MM-Debiased dataset. Other methods: OGM-GE [21], Greedy [33] and PMR [4].

As deep learning improves by leaps and bounds, deep neural networks with the capacity to learn representation from a large amount of data have been used extensively in multimodal learning [31]. Although the deep-based methods have revealed effectiveness, recent studies have found the imbalanced multimodal learning problem where most existing models often have unsatisfactory multimodal cooperation, which cannot jointly utilize all modalities well [11, 33]. But deep models' lack of interpretability makes it hard to observe what role each modality plays in the final prediction, and then accordingly adjust the uni-modal training. Some methods have been proposed to identify and improve the training of worse learnt modality with the help of output logits or the scale of gradient [21, 33]. These empirical strategies only consider the global modality discrepancy at dataset-level, and achieve improvement on the common curated dataset (as Kinetics Sounds dataset in Figure 1). *However, under realistic scenarios, the modality discrepancy could vary across different samples.* For example, Figure 2a and 2b show two audio-visual samples of *motorcycling* category. The motorcycle in *Sample 1* is hard to

[*]Corresponding author.



(a) Visual of *S.1* of *motorcycling*. (b) Visual of *S.2* of *motorcycling*. (c) Valuation of *S.1* and *S.2*. (d) Avg. Contribution of dataset.

Figure 2. **(a-b):** Audio-visual samples of *motorcycling* category. **(c):** Our modality valuation of *S.1* and *S.2*. *S.1* and *S.2* denotes *Sample 1* and *Sample 2* respectively. **(d):** Uni-modal average contribution over all training samples of different dataset. Our proposed MM-Debiased dataset has less global discrepancy at dataset-level, compared with other curated dataset.

observe while the wheel of motorcycle in *Sample 2* is quite clear. This could make audio or visual modality contribute more to the final prediction respectively for these two samples. This fine-grained modality discrepancy is hard to perceive by existing methods. Hence, how to *reasonably* observe and improve multimodal cooperation at sample-level is still expected to be resolved.

To this end, we introduce a sample-level modality valuation metric, to observe the contribution of each modality during prediction for each sample. The Shapley value from game theory [23], which aims to fairly distribute the benefits based on the contribution of each player, provides the theoretical support of our valuation. Via valuating uni-modal contribution, we observe that the experiment results unsatisfactorily fail to meet the expectation that each modality has its irreplaceable contribution. Firstly, as Figure 2d, for existing curated dataset including Kinetics Sounds and UCF-101, we verify that the contribution of one modality tends to overwhelm others globally at the dataset-level. More importantly, as Figure 2c, with our sample-level modality valuation, we observe that modality discrepancy indeed could be different across samples, beyond the global contribution discrepancy at dataset-level. To highlight this sample-level modality discrepancy, we propose the global balanced MM-Debiased dataset where the dataset-level modality discrepancy is no longer significant (as Figure 2d). Not surprisingly, existing imbalanced multimodal learning methods which only consider dataset-level discrepancy *fail* on MM-Debiased dataset, as shown in Figure 1.

Based on the above empirical results, we first analyze the effect of the modality with clearly lower contribution in a sample and find that *its presence would potentially increase the risk that the multimodal model collapses to one specific modality*. Hence, it is urgent to recover the suppressed contribution of these low-contributing modalities. To alleviate the above problem, we further analyze the correlation between uni-modal discriminative ability and its contribution, then find that *enhancing the discriminative ability of low-contributing modality during training could indirectly improve its contribution in a sample, and accordingly enhance multimodal cooperation*. Therefore, we propose to train the low-contributing modality in a sample in a targeted manner based on the contribution discrepancy between modalities. Specifically, we first valuate the uni-modal contribution at sample-level via our Shapley-based modality valuation metric. Then the input of identified low-contributing modalities is re-sampled with a dynamical frequency, determined by the exact contribution discrepancy, to improve its discriminative ability targetedly. Considering the computational cost of sample-wise modality valuation, we also propose the more efficient modality-level method.

As Figure 1, our methods considering the sample-level modality discrepancy achieve considerable improvement on both existing curated and our global balanced dataset, Overall, our contributions are as follows. **Firstly,** we introduce a sample-level modality valuation metric and further analyze the low-contributing modality issue, which could worsen the multimodal cooperation. **Secondly**, methods are proposed to strengthen low-contributing modalities, reasonably improving multimodal cooperation. **Thirdly**, we propose the MM-Debiased dataset with fine-grained multimodal discrepancy, which is closer to the realistic scenario.

2. Related works

Imbalanced multimodal learning. Recent studies have found the multimodal model has a preference for specific modality [11]. Several methods have been proposed to ease this problem by improving the optimization of worse learnt modality [4, 21, 33–35]. These methods often control uni-modal optimization by estimating the discrepancy of the training stage or performance between modalities. However, their estimation could be hard to observe modality discrepancy at sample-level, handling realistic scenarios where the performance difference of each modality could vary across samples. In this paper, we go a step further, reasonably valuating the uni-modal contribution at sample-level using the introduced Shapley-based metric. This fine-grained modality valuation metric could guide us to solve the imbalanced multimodal learning problem better.

Game theory in machine learning. Researchers have adapted the game theory to formulate and solve machine learning problems [5, 6, 29]. For example, game theory has been used to explain the algorithm effectiveness of AdaBoost [5]. Similarly to us, Hu et al. [9] use the Shapley value to evaluate the overall contribution of individual modalities for the whole dataset. But they cannot capture the modality contribution at sample-level and do not provide further analysis or methods. In this paper, we not only introduce sample-level modality valuation, but also further analyze and alleviate the low-contributing modality issue.

3. Method

3.1. Model formulation

In this paper, we consider the multimodal discriminative task. Concretely, each sample $x = (x^1, x^2, \cdots, x^n)$ is with n modalities. And y is the ground truth label of sample x. For simplicity, the input of modality i of specific sample x is denoted as x^i. $N = \{x^1, x^2, \cdots, x^n\}$ is a finite, nonempty set of all modalities. Denote the multimodal model as $H(\cdot)$. Suppose C is the set of input modalities for the model, where $C \subseteq N$. When taking modalities in C as the input, the final prediction is $\widehat{y_C} = H(\cup x^i, x^i \in C)$. It should be noted that we have no assumption of multimodal fusion design, therefore the following modality valuation is not limited to simple fusion strategy.

3.2. Fine-grained modality valuation

In multimodal learning, each modality is expected to fully demonstrate its irreplaceable contribution, since different modalities are considered with complementary information. Based on realistic scenarios, the modality contribution discrepancy could vary across different samples. Hence, it is necessary to valuate the uni-modal contribution in the multimodal model at sample-level, and accordingly improve multimodal cooperation. In this paper, we introduce a Shapely-based metric fine-grained modality valuation metric, to observe the uni-modal contribution for the multimodal prediction at sample-level.

Concretely, for each sample x, we first have v as a function to map the multimodal prediction to its benefits:

$$v(C) = \begin{cases} |C| & \text{if } \widehat{y_C} = y, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

When predicting correctly, the benefits of multimodal prediction with input C is the number of input modalities.

After formulating the prediction benefits, to consider the contribution of each modality under all cases, let Π_N denote the set of all permutations of N. Since the number of modalities is n, there is $|\Pi_N| = n!$. For modality i of sample x, given a permutation $\pi \in \Pi_N$, we denote by $S_\pi(x^i)$ the set of all predecessors of it in π, *i.e.,* we set $S_\pi(x^i) = \{x^j \in N | \pi(x^j) < \pi(x^i)\}$. The *marginal contribution* of modality i of sample x with respect to a permutation π is denoted by $\Delta_\pi(x^i)$ and is given by:

$$\Delta_\pi(x^i) = v(S_\pi(x^i) \cup x^i) - v(S_\pi(x^i)). \tag{2}$$

This quantity measures how much modality i increases the benefits of its predecessors in π when it joins the permutation π. Since different combinations of modalities could have different results, we should consider all possible permutations to fully evaluate uni-modal contribution. Then, given Π_N with $n!$ permutations, the *final contribution* of modality i is denoted by ϕ^i and is given by:

$$\phi^i = \frac{1}{n!} \sum_{\pi \in \Pi_N} \Delta_\pi(x^i). \tag{3}$$

It should be noted that when considering all possible permutations, the sum of all uni-modal contribution $\sum_{i=1}^n \phi^i$ is in fact the benefits of multimodal prediction with all modalities as the input. Hence, for the normal model with all modalities as the input, when the contribution of one modality increases, the contribution of other modalities would accordingly decrease. With this sample-level modality valuation metric, we could reasonably observe the uni-modal contribution for each sample.

3.3. Low-contributing modality phenomenon

As Figure 2, both at sample-level and dataset-level, contribution of one modality could highly overwhelm others. In other words, the decision of multimodal model is dominated by one modality, remaining others low-contributing.

Here we analyze the effect of low-contributing modality to the benefits of normal multimodal model for sample x. Suppose the marginal contribution of modality in multimodal learning is non-negative, since the introduction of additional modality tends to not bring negative effects in practice. Based on the definition of uni-modal contribution for modality i, we have:

$$\phi^i = \frac{1}{n!} \sum_{\pi \in \Pi_N} \Delta_\pi(x^i), \tag{4}$$

$$\phi^i = \frac{1}{n!} \sum_{\pi \in \Pi_N} (v(S_\pi(x^i) \cup x^i) - v(S_\pi(x^i))), \tag{5}$$

$$n! \cdot \phi^i \geq \underbrace{(n-1)! \cdot (v(N) - v(N \backslash x^i))}_{\text{only consider cases } x^i \text{ is the last one, which have } (n-1)! \text{ permutations}}, \tag{6}$$

$$n \cdot \phi^i \geq v(N) - v(N \backslash x^i), \tag{7}$$

$$v(N) - v(N \backslash x^i) \leq n \cdot \phi^i. \tag{8}$$

In addition, based on Equation 1, when predicting correctly, $v(N) = n$. We know that the minimum of $v(N \backslash x^i)$ is 0. Then we have:

$$v(N) - v(N \backslash x^i) \leq n. \tag{9}$$

However, based on Equation 8, when $\phi^i < 1$, the upper bound of the difference between $v(N)$ and $v(N \backslash x^i)$ shrinks (*i.e.,* $n \cdot \phi^i < n$). In other words, when the contribution of modality i, $\phi^i < 1$, the difference between $v(N)$ and $v(N \backslash x^i)$ decreases. The benefits of taking all modalities N as the input becomes close to its subset $N \backslash x^i$. Assuming an extreme case where the contribution of all but one modality is very small, multimodal learning is close to uni-modal learning. Hence, it is essential to enhance the contribution of low-contributing modalities for each sample, improving multimodal cooperation.

Remark 1. *Suppose the marginal contribution of modality is non-negative. For the normal multimodal model with all modalities of sample x as the input, with benefits $v(N) = n$, when modality i is low-contributing,* i.e., *$\phi^i < 1$, the difference between $v(N)$ and $v(N \backslash x^i)$ decreases.*

In Remark 1, we suppose the marginal contribution of modality is non-negative. In practice, the introduction of additional modalities has been validated its benefit (non-negative effect) across different application tasks [16]. It also theoretically proves that multimodal learning provably performs better than uni-modal [10]. These evidences indicate that the introduction of another related modality could not bring a negative impact in most cases. Based on this, we assume that the marginal contribution is non-negative.

To alleviate the above problem, we further analyze the correlation between uni-modal discriminative ability and its contribution and have Remark 2. Based on the analysis, strengthening the discriminative ability of low-contributing modality can improve its contribution to multimodal prediction. Correspondingly, the risk that multimodal model collapses to one specific modality is lowered[1].

Remark 2. *Suppose the marginal contribution of modality is non-negative and the numerical benefits of one modality's marginal contribution follow the discrete uniform distribution. Enhancing the discriminative ability of low-contributing modality i can increase its contribution ϕ^i.*

3.4. Re-sample enhancement strategy

Based on Remark 2, enhancing the discriminative ability of low-contributing modality can expand its contribution. Hence, we propose to improve the discriminative ability of low-contributing modality during training by re-sampling its input in a targeted manner.

Concretely, to ensure the basic discriminative ability, we first warm up the multimodal model for several epochs. Then, at each epoch, modality valuation is conducted once to observe uni-modal contribution for each sample. Subsequently, learning of the low-contributing modality can be targetedly improved via solely re-sampling its input. Here, we provide the fine-grained as well as effective sample-level re-sample method and the coarse but efficient modality-level re-sample method.

Algorithm 1 Sample-level method

```
Require: Original training dataset $\mathcal{D}$, training dataset with
  re-sample $\mathcal{D}^{\text{rs}}$, number of modalities $n$, model parameters
  $\theta$, training epoch $T$, warm-up epoch $F$.
  for $t = 0, \cdots, T-1$ do
    if $t < F$ then
      Update model parameters $\theta$ with dataset $\mathcal{D}$;
    else
      Initialize $\mathcal{D}^{\text{rs}}$: $\mathcal{D}^{\text{rs}} = \mathcal{D}$;
      for each sample $x$ in $\mathcal{D}$ do
        Obtain uni-modal contribution $\{\phi^1, \phi^2, \cdots, \phi^n\}$
        with Equation 3;
        Identify modality $i$ where $\phi^i < 1$;
        Get frequency $s(x^i)$ with Equation 10;
        Add $x^i$ with frequency $s(x^i)$ into $\mathcal{D}^{\text{rs}}$;
      end for
      Update model parameters $\theta$ with dataset $\mathcal{D}^{\text{rs}}$;
    end if
  end for
```

3.4.1 Sample-level method

After the modality valuation, the low-contributing modality i, $\phi^i < 1$, for each sample, can be well distinguished and we can finely improve its learning at sample-level. Then the specific re-sample frequency is dynamically determined by the exact value of ϕ^i during training. Specifically, the re-sample frequency of modality i for specific sample x is:

$$s(x^i) = \begin{cases} f_s(1-\phi^i) & \phi^i < 1, \\ 0 & \text{others}, \end{cases} \tag{10}$$

where $f_s(\cdot)$ is a monotonically increasing function. Utilizing this sample-level re-sample strategy, the low-contributing modality i in sample x is re-trained with a re-sample frequency inversely proportional to its contribution, *i.e.,* the less the contribution is, the larger the re-sample frequency is. It is worth noting that only the low-contributing modality is taken during re-sampling, and inputs of other modalities are masked by 0, to ensure targeted learning.

3.4.2 Modality-level method

Although sample-level modality valuation could provide fine-grained uni-modal contribution, there could be a high additional computational cost when the scale of dataset is quite large. Therefore, the more efficient modality-level method is proposed to lower cost. As Figure 2d, the low-contributing phenomenon has a dataset-level preference. For example, the average contribution of RGB modality

[1] The specific theoretical analysis process of Remark 2 is provided in the *Supp. Materials*.

over all training samples is obviously more than that of optical flow modality on the UCF-101 dataset. Hence, we propose a more coarse modality-level re-sample strategy, which estimates the average uni-modal contribution via only conducting modality valuation on the subset of training samples to reduce additional computational cost.

Concretely, we randomly split a subset with Z samples in the training set to approximately estimate the average uni-modal contribution. Hence, the overall low-contributing modality i with less average ϕ^i, *i.e.,* $\frac{\sum_{k=1}^{Z}\phi_k^i}{Z}$, can be identified. Then, other modalities remain unchanged, and modality i in sample x is dynamically re-sampled with specific probability $p(i)$ during training:

$$p(i) = f_m(\text{Norm}(d)), \tag{11}$$

where $d = \frac{1}{n-1}(\sum_{j=1,j\neq i}^{n}(\frac{\sum_{k=1}^{Z}\phi_k^j}{Z} - \frac{\sum_{k=1}^{Z}\phi_k^i}{Z}))$. The discrepancy in average contribution between overall low-contributing modality i compared to other modalities (*i.e., d*) is first $0-1$ normalized, then fed into $f_m(\cdot)$, a monotonically increasing function with a value between 0 and 1. This re-sample probability for overall low-contributing modality is proportional to the discrepancy in its average contribution compared to others. Compared to sample-level strategy, modality-level one is more efficient.

Algorithm 2 Modality-level method

```
Require: Original training dataset $\mathcal{D}$, training dataset with
    re-sample $\mathcal{D}^{\text{rs}}$, subset of training dataset $Z$, number of
    modalities $n$, model parameters $\theta$, training epoch $T$,
    warm-up epoch $F$.
  for $t = 0, \cdots, T-1$ do
    if $t < F$ then
      Update model parameters $\theta$ with dataset $\mathcal{D}$;
    else
      Initialize $\mathcal{D}^{\text{rs}}$: $\mathcal{D}^{\text{rs}} = \mathcal{D}$;
      for each sample $x$ in $Z$ do
        Obtain uni-modal contribution $\{\phi^1, \phi^2, \cdots, \phi^n\}$
        with Equation 3;
      end for
      Identify overall low-contributing modality $i$;
      Get re-sample probability $p(i)$ with Equation 11;
      for each sample $x$ in $\mathcal{D}$ do
        Add $x^i$ with probability $p(i)$ into $\mathcal{D}^{\text{rs}}$;
      end for
      Update model parameters $\theta$ with dataset $\mathcal{D}^{\text{rs}}$;
    end if
  end for
```

4. Experiment

4.1. Dataset and experimental settings

Kinetic Sounds (KS) [1] is an action recognition dataset with two modalities, audio and video. This dataset contains 31 human action classes, which are selected from Kinetics dataset [13]. It contains 19k 10-second video clips.

UCF-101 [26] is an action recognition dataset with two modalities, RGB and optical flow. This dataset contains 101 categories of human actions. The entire dataset is divided into a 9,537-sample training set and a 3,783-sample test set according to the original setting.

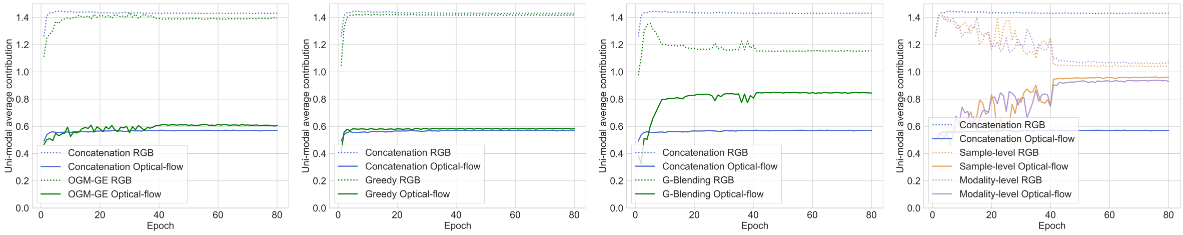
MM-Debiased is an audio-visual dataset proposed by us, where dataset-level modality contribution discrepancy is not obvious. It covers 10 classes, and contains 11,368 training samples and 1,472 testing samples. Details about the dataset construction are provided in *Supp. Materials*.

Experimental settings. When not specified, ResNet-18 [8] is used as the backbone in experiments. Encoders used for UCF-101 are ImageNet pre-trained. Encoders of other datasets are trained from scratch. During the training, we use SGD with momentum (0.9) and set the learning rate at $1e-3$. A subset with 20% training samples is randomly split in modality-level method. During modality valuation, for input modality set C, input of modalities not in C are zeroed out, similar to related work [7]. During testing, all modalities are taken as the model input. Detailed experimental settings, experiments about more than two modalities, and ablation studies about the subset scale, $f_s(\cdot)$ as well as $f_m(\cdot)$, are provided in *Supp. Materials*.

4.2. Comparison with multimodal fusion methods

Here we first compare our methods with several representative multimodal fusion methods under deep frameworks: Concatenation [20], Summation, Decision fusion [25], FiLM [22] and Gated [14]. Besides, the early multimodal integration attempts, Bayesian network [2] and Multi-kernel Learning (MKL) [24], are also compared. To be fair, the uni-modal encoders of Bayesian network are ResNet-18 and features fed into MKL are extracted by pre-trained uni-modal encoders. Our sample-level and modality-level methods are based on Concatenation fusion in Table 1.

Based on Table 1, several observations can be revealed. Firstly, early multimodal integration methods are able to disclose their effectiveness after being equipped with extracted deep features, especially MKL, which even outperforms the Concatenation model. However, this improvement has a reliance on the quality of input features, and these methods are still hard to directly process raw data in large-scale. Secondly, both our sample-level and modality-level strategies improve multimodal cooperation by means of fine-grained modality valuation, achieving better model performance. Moreover, the fine-grained sample-level method tends to be superior. In contrast, the modality-level method is more efficient and sometimes even can be comparable to sample-level one.



(a) OGM-GE [21]. (b) Greedy [33]. (c) G-Blending [30]. (d) Ours.

Figure 3. Average contribution of each modality over all training samples during training for OGM-GE, Greedy, G-Blending and our methods on the UCF-101 dataset.

Method	KS (Audio+Visual)		UCF-101 (RGB+OF)	
	Acc	mAP	Acc	mAP
Concatenation	62.30	67.95	81.15	86.15
Summation	62.10	66.97	81.31	86.25
Decision fusion	62.65	67.89	79.81	86.07
FiLM [22]	61.25	64.85	79.45	84.27
Gated [14]	62.72	68.28	81.34	86.35
Bayesian DNN [2]	60.79	64.98	80.04	84.95
Deep MKL* [24]	63.61	69.69	82.64	87.96
Sample-level	**66.92**	<u>71.84</u>	**83.52**	**88.89**
Modality-level	<u>66.65</u>	**72.68**	<u>83.46</u>	<u>88.75</u>

Table 1. **Comparison with different multimodal fusion methods.** Bold and underline represent the best and runner-up respectively. * denotes that the fed feature of Deep MKL model is extracted by pre-trained uni-modal encoders. Bayesian DNN is trained from scratch. OF denotes for optical flow. *Due to limited space, experiments about more modalities, e.g., text modality, are provided in Supp. Materials.*

4.3. Comparison with imbalanced multimodal learning methods

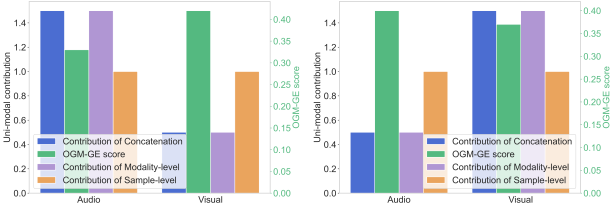
Recent studies have found that multimodal models often cannot jointly utilize all modalities well, and some imbalanced multimodal learning methods are proposed. They often control uni-modal optimization by estimating the discrepancy of the training stage or performance between modalities. Here we compare with these imbalanced multimodal learning methods, OGM-GE [21], G-Blending [30], Greedy [33], PMR [4] and AGM [15]. Our sample-level and modality-level methods are based on Concatenation fusion.

Common case. In many dataset, like Kinetics Sounds and UCF-101, as the former analysis, the low-contributing phenomenon has a dataset-level preference. Firstly, based on the results in Table 2, in this case with dataset-level modality preference, these methods often have a performance improvement. Among them, our methods outperform these imbalanced multimodal learning approaches. Although G-Blending [30] achieves considerable performance, it needs to train an additional uni-modal classifier as the basis of modulation. Concretely, in fact, FLOPs of our methods reduce $1/4$ (sample-level method), even $1/2$ (modality-level method), compared to G-Blending. Secondly, since our modality valuation is not limited to specific methods, the uni-modal contribution of other methods can also be observed. As Figure 3, our methods exhibit superior mitigation of imbalanced uni-modal contributions, thereby highlighting our efficacy beyond mere final prediction.

Method	KS	UCF-101	MM-Debiased
Concatenation	62.30	81.15	83.22
OGM-GE [21]	63.30	81.54	83.08 (↓)
G-Blending [30]	66.24	83.09	84.17
Greedy [33]	62.37	81.25	83.08 (↓)
PMR [4]	63.23	81.04 (↓)	82.88 (↓)
AGM [15]	63.96	81.65	83.22 (−)
Sample-level	**66.92**	**83.52**	**84.71**
Modality-level	<u>66.65</u>	<u>83.46</u>	<u>84.31</u>

Table 2. Accuracy of imbalanced multimodal learning methods, where bold and underline represent the best and runner-up respectively. ↓ indicates a performance drop compared with Concatenation baseline.

More balanced case. However, under realistic scenarios, the modality discrepancy could vary across different samples, as samples shown in Figure 2a and 2b. To comprehensively evaluate imbalanced multimodal learning methods, we conduct experiments under the more balanced case, where dataset-level discrepancy is not apparent, but there are still sample-level modality discrepancy. We build the MM-Debiased dataset where the dataset-level discrepancy is no longer significant. As Figure 2d, the average contribution of each modality on the MM-Debiased dataset is more balanced than that on other curated dataset. Based on the results shown in Table 2, most imbalanced multimodal learning methods including OGM, Greedy and PMR are even worse than Concatenation baseline, since existing methods only consider the dataset-level modality preference. They could not capture the sample-level modality discrepancy. In contrast, our methods, especially our sample-level method, achieves considerable improvements. Our method can reasonably valuate fine-grained modality contribution, and targetedly enhance the learning of low-contributing modality.



(a) Valuation of *Sample 1*. (b) Valuation of *Sample 2*.

Figure 4. Valuation of two samples of *motorcycling* category.

4.4. Comparison of sample-level modality valuation

Beyond the model performance, we further conduct experiments to compare with existing imbalanced multimodal learning methods about sample-level modality valuation. In these methods, to modulate the uni-modal optimization, they also evaluate specific uni-modal properties. For example, G-Blending [30] and Greedy [33] inspect the uni-modal training process. AGM [15] proposes to evaluate the modality contribution which is then used to modulate gradient. But they could not be used to evaluate the modal preference at the sample-level. The uni-modal confidence score used by OGM-GE [21] could be used to evaluate uni-modal performance at sample-level. However, this empirically designed score is hard to handle in realistic scenarios, like dominant modality could differ among samples within the same category, since its calculation could suffer from the dataset-level discrepancy, resulting in inaccurate results.

For example, for the two audio-visual samples of *motorcycling* category in Figure 2a and 2b, the motorcycle in *Sample 1* appears hard to be observed, while the wheel of motorcycle in *Sample 2* appears clearly. These two samples receptively rely on audio or visual modality. Here we propose the modality valuation of different methods of these two samples. Results are shown in Figure 4. Based on the contribution of Concatenation baseline produced by our methods (blue bar in Figure 4), our valuation correctly reflects that these two samples rely on audio and visual signals, respectively. However, OGM-GE score provides the wrong results, assigning more confidence for the less informative visual signal of *Sample 1* (green bar in Figure 4a).

Moreover, our fine-grained sample-level method captures and accordingly adjusts the uni-modal learning, balancing this fine-grained modality discrepancy (orange bar in Figure 4). Although our modality-level method fails to ease this discrepancy (purple bar in Figure 4), it has an advantage in efficiency. These experiments also indicate that the sample-level and modality-level methods have their own advantages and applicable scenarios.

4.5. Complex cross-modal interaction scenarios

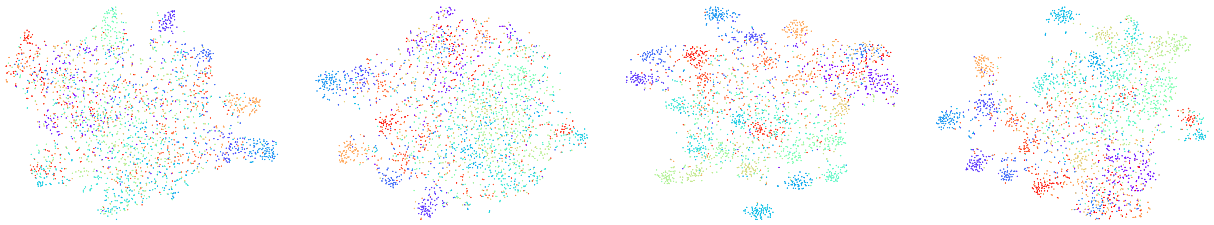
As stated before, different from most existing imbalanced multimodal learning methods, our methods are not limited to simple fusion strategies. Here we first combine our sample-level and modality-level methods with two intermediate fusion methods, CentralNet [28] and MMTM [12], to evaluate their effectiveness under these cross-modal interaction scenarios. Based on the results in Table 3, these cross-modal interaction modules indeed improve the model performance, compared with Concatenation baseline. This phenomenon demonstrates that the cross-modal interaction could implicitly deepen the cooperation between modalities by helping one modality make adjustments according to the feedback from others.

Method	KS	UCF-101	MM-Debiased
Concatenation	62.30	81.15	83.22
Concat-Sample	**66.92**	**84.71**	**84.31**
Concat-Modality	66.65	83.46	84.04
CentralNet [28]	67.35	83.97	85.23
CentralNet-Sample	67.89	**84.07**	**85.39**
CentralNet-Modality	**68.31**	84.05	85.26
MMTM [12]	64.23	80.67	84.31
MMTM-Sample	**64.40**	**81.30**	**85.33**
MMTM-Modality	64.34	81.23	84.71
MBT [17]	47.02	-	68.01
MBT-Sample	**47.53**	-	**68.70**
MBT-Modality	47.36	-	68.34

Table 3. **Accuracy of multimodal frameworks with cross-modal interaction modules.** Results of MBT on UCF-101 dataset could not be obtained since samples of it is hard to be trained from scratch but lacking suitable pre-trained transformer backbone.

In addition, our methods can be well applied to these more complex scenarios with cross-modal interaction, bringing performance improvement. One additional observation is that although the architecture limits model performance, our methods applied to the simple Concatenation fusion method could even have comparable results with more complex model designs, indicating it is simple-yet-effective. Furthermore, to qualitatively observe the quality of uni-modal representations, we visualize the feature distribution of overall low-contributing modality encoder on the Kinetics Sounds dataset (*i.e.,* the visual modality). Results in Figure 5 illustrate that the feature distribution is more discriminative in terms of action categories with cross-modal interaction, and this discriminative distribution can go a step further after being equipped with our methods.

Besides these modules based on CNN backbone, transformer networks also have cross-modal interaction. Here we combine our methods with the representative transformer model, MBT [17], to explore their effectiveness. Results are in Table 3. Models are trained from scratch. It should be noted that the performance of MBT is inferior to Concatenation with CNN backbone, since the transformer network is generally data-hungry, limiting its performance on these datasets without large enough samples,

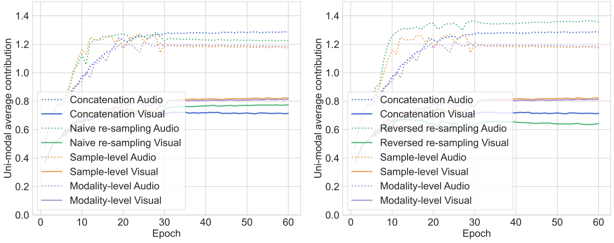


(a) Concatenation. (b) MMTM [12]. (c) MMTM-Sample. (d) MMTM-Modality.

Figure 5. Visual feature distribution of Concatenation, MMTM, MMTM-Sample and MMTM-Modality, visualized by t-SNE [27] on Kinetics Sounds dataset. As Figure 2d, visual modality tends to be the low-contributing one. Categories are indicated in different colors.

Method	**KS**		**UCF-101**	
	Acc	**mAP**	**Acc**	**mAP**
Concatenation	62.30	67.95	81.15	86.15
Naïve re-sample	64.69	70.87	82.56	88.02
Reversed re-sample	59.03	63.14	80.85	85.06
Sample-level	**66.92**	71.84	**83.52**	**88.89**
Modality-level	66.65	**72.68**	83.46	88.75

Table 4. **Comparison with other re-sample strategies.** Bold and underline represent the best and runner-up respectively.



(a) Naïve comparison. (b) Reversed comparison.

Figure 6. Average contribution of each modality over all training samples during training for Naïve re-sample and Reversed re-sample methods on the Kinetics Sounds dataset.

when being trained from scratch. But both our sample-level and modality-level methods can combine with and further enhance the performance of MBT. Overall, our methods could be well equipped with different cross-modal interaction modules, bringing performance enhancement.

4.6. Comparison with other re-sample strategies

To further validate our methods under the guidance of sample-level modality valuation, we compare with two related re-sample settings, naïve re-sample and reversed re-sample. Naïve re-sample is to randomly re-sample input of each modality with the same frequency as ours, while reversed re-sample is contrary to our methods, only re-sampling the data of modality with higher contribution.

Based on Table 4, naïve re-sample method can also increase model ability. The reason could be that this random uni-modal re-sample setting also potentially provides the chance for each modality to be trained individually, improving the discriminative ability of low-contributing modality. Hence, its contribution is accordingly boosted based on Remark 2. As the average uni-modal contribution shown in Figure 6a, naïve re-sample method indeed alleviates the low-contributing issues to some extent (the green line). Beyond that, our methods with targeted re-sample design under the guidance of sample-level modality valuation during training, take one step further (as the orange and purple lines). In addition, the failure of reversed re-sample setting (the green line in Figure 6b), which runs counter to our analysis, also validates that it is our modality valuation guidance that matters, rather than the re-sample strategy itself.

5. Discussion

In this paper, we introduce a sample-level modality valuation metric to observe uni-modal contribution with the aid of theory in game theory. Two methods are proposed to recover the suppressed contribution of low-contributing modality, improving multimodal cooperation. Besides, there is also some further discussion.

Natural difference between modalities. In practice, there is a natural difference between modalities. For example, for an audio-visual sample *drawing picture*, vision is naturally more discriminative than auditory. Hence, our methods could recover the suppressed contribution of low-contributing modality, but could not make the uni-modal contribution equal. Therefore, it is expected to evaluate and take this natural difference into consideration during improving multimodal cooperation in the future.

Imbalanced contribution in Multimodal Large Language Model. Recently, the development of Multimodal Large Language Model has widely spread attention. However, in these models, like GPT-4V, there is also an imbalanced contribution issue. For example, results of GPT-4V are more likely to be misled by text modality [3]. To this end, the study about imbalanced uni-modal contribution is expected to extend to this case.

Acknowledgements. This research was supported by National Natural Science Foundation of China (NO.62106272).

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 5

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 5, 6

[3] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023. 8

[4] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 1, 2, 6

[5] Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 325–332, 1996. 2, 3

[6] Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. Eigengame: Pca as a nash equilibrium. In *International Conference on Learning Representations*, 2020. 2

[7] Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. *Advances in Neural Information Processing Systems*, 33:5922–5932, 2020. 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[9] Pengbo Hu, Xingyu Li, and Yi Zhou. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*, 2022. 3

[10] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021. 4

[11] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). *arXiv preprint arXiv:2203.12221*, 2022. 1, 2

[12] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020. 7, 8

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[14] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5, 6

[15] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023. 6, 7

[16] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022. 4

[17] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 7

[18] Ara V Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1–15, 2002. 1

[19] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 1

[20] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 5

[21] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 1, 2, 6, 7

[22] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5, 6

[23] Lloyd S. Shapley. 17. a value for n-person games. 1953. 2

[24] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524, 2013. 5, 6

[25] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 5

[26] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[27] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[28] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 7

[29] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks:

introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017. 2

[30] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 6, 7

[31] Yang Wang. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–25, 2021. 1

[32] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022. 1

[33] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022. 1, 2, 6, 7

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

[35] Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. In *The Twelfth International Conference on Learning Representations*, 2023. 2