On-the-fly Modulation for Balanced Multimodal Learning

Yake Wei, Di Hu, Henghui Du, Ji-Rong Wen, *Senior Member, IEEE*

Abstract—Multimodal learning is expected to boost model performance by integrating information from different modalities. However, its potential is not fully exploited because the widely-used joint training strategy, which has a uniform objective for all modalities, leads to imbalanced and under-optimized uni-modal representations. Specifically, we point out that there often exists modality with more discriminative information, *e.g.,* vision of *playing football* and sound of *blowing wind*. They could dominate the joint training process, resulting in other modalities being significantly under-optimized. To alleviate this problem, we first analyze the under-optimized phenomenon from both the feed-forward and the back-propagation stages during optimization. Then, On-the-fly Prediction Modulation (OPM) and On-the-fly Gradient Modulation (OGM) strategies are proposed to modulate the optimization of each modality, by monitoring the discriminative discrepancy between modalities during training. Concretely, OPM weakens the influence of the dominant modality by dropping its feature with dynamical probability in the feed-forward stage, while OGM mitigates its gradient in the back-propagation stage. In experiments, our methods demonstrate considerable improvement across a variety of multimodal tasks. These simple yet effective strategies not only enhance performance in vanilla and task-oriented multimodal models, but also in more complex multimodal tasks, showcasing their effectiveness and flexibility. The source code is available at https://github.com/GeWu-Lab/BML_TPAMI2024.

Index Terms—Multimodal learning, On-the-fly Prediction Modulation, On-the-fly Gradient Modulation

1 Introduction

People perceive the surrounding world by comprehensively integrating multiple senses, including vision, hearing, and touch. This process is known in cognitive neuroscience as multi-sensory integration [1]. Inspired by this phenomenon, multimodal data, collected from multiple sensors, has raised attention in the machine learning field, and accordingly multimodal learning has witnessed significant advances in these years. The research community has improved the performance of traditional uni-modal tasks by incorporating additional modalities and has also begun tackling new, challenging problems [2], such as multimodal action recognition [3], [4], multimodal semantic segmentation [5], [6] and audio-visual event localization [7].

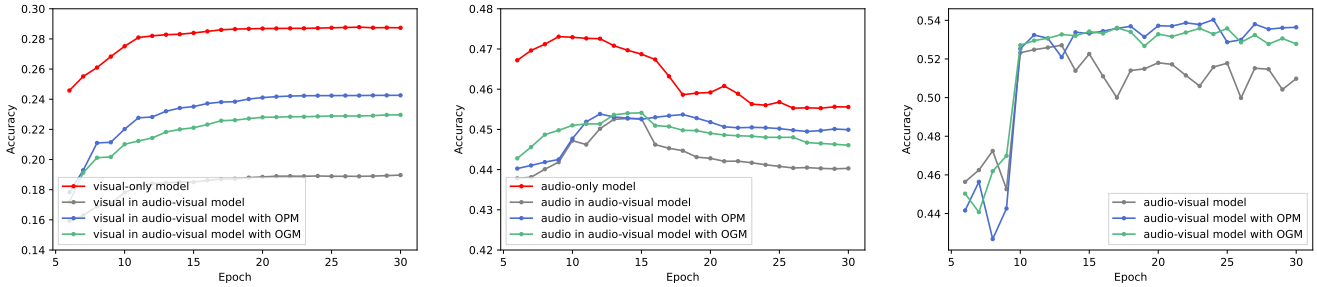
Multimodal models are expected to surpass their uni-modal counterparts since they take data containing information from multiple views. In most cases, it does achieve this intention, but sometimes goes the contrary: the multimodal model can be inferior to the uni-modal one [8]. In recent studies, some researchers claimed that different modalities could perform dis-similarly in the optimization process. For instance, the audio modality tends to converge with a faster learning pace in the video recognition task, compared with the visual one [8]. This discrepancy makes it challenging for multimodal models to effectively learn from all modalities simultaneously under a uniform joint training objective [9], [10]. As a result, the potential of multimodal models can be limited by the difference in the learning status of different modalities, and then fail to outperform uni-modal counterparts. To cope with this issue, some studies depending on added uni-modal classifiers or additional training for the specific modality are proposed [8], [11], but they inevitably introduce extra training efforts.

Beyond the failure cases of multimodal joint learning, we note that even when multimodal models outperform their uni-modal counterparts, they still fail to fully harness the potential of multiple modalities. As demonstrated in Fig 1, we conducted experiments on the VGGSound dataset [13] to assess the quality of uni-modal encoders in jointly trained multimodal models. Our results show that the jointly trained multimodal models perform better than the uni-modal models, which is expected. However, when examining the performance of uni-modal encoders within these multimodal models[1], we discover that they are under-optimized compared to the corresponding solely trained uni-modal models. For example, in Fig 1(a), during the whole training process, performance of visual-only model (red line) is better than visual encoder in audio-visual model (gray line). Moreover, *the under-optimized degrees of different modalities are imbalanced, and one modality is clearly worse learnt than others*. As shown in Fig 1(a) and Fig 1(b), the quality of visual encoder in the audio-visual model has a more clear drop, compared with the audio modality. Overall, these interesting observations demonstrate that the uni-modal representation is under-optimized with an imbalanced degree in the joint training multimodal model.

The reason could be that, for the multimodal dataset,

• *Y. Wei, D. Hu (corresponding author), H. Du, and J-R. Wen are with the Gaoling School of Artificial Intelligence, and Beijing Key Laboratory of Big Data Management and Analysis Methods, Renmin University of China, Beijing 100872, China.*
E-mail: {yakewei, dihu, cserdu, jrwen}@ruc.edu.cn

1. Here the audio-only and visual-only models are trained individually. To evaluate the quality of uni-modal encoder in jointly trained audio-visual models at a certain epoch, we first take the uni-modal encoder in audio-visual models at this epoch, then freeze its parameters and fine-tune a new uni-modal classifier. Finally, the accuracy of the uni-modal encoder in audio-visual models at a certain epoch is obtained.



(a) (b) (c)

Fig. 1. **Performance of individually trained uni-modal model, jointly trained multimodal model and jointly trained multimodal model with our proposed OPM and OGM strategies respectively on the VGGSound dataset.** (a) Performance of visual modality. (b) Performance of audio modality. (c) Performance of audio-visual modalities. Best viewed in color. The training of our OPM and OGM methods exactly aligns with the applied audio-visual model. To provide more representative observation, here jointly trained multimodal use concatenation fusion, which is widely-used, and simple-but-strong. In Appendix D, we also extend these experiments to more complex CentralNet [12] multimodal framework.

there often exists a dominant modality [14] with the better discriminative ability (*e.g.*, vision of *drawing* and sound of *wind blowing*), which tends to be favored during training and consequently suppress the learning of others. For instance, as illustrated in Fig 1, the visual encoder in the audio-visual model is more remarkably under-optimized. This observation is consistent with the fact that the curated sound-oriented dataset, VGGSound, has a preference for the audio modality. This preference within the dataset would result in one modality often being more discriminative, causing the observed imbalanced learning problem among modalities.

To alleviate this problem, we first analyze the imbalanced learning phenomenon in the jointly trained multimodal model from both the feed-forward and back-propagation stages. In the feed-forward stage, the modality with more discriminative information often determines the model output and dominates the prediction. Subsequently, it also lowers the joint loss more, limiting the gradient of other modalities in the back-propagation stage. These observations cause the imbalanced learnt situation between modalities. To ease this issue, we propose to control the optimization of each modality via two on-the-fly modulation methods: *On-the-fly Prediction Modulation* (OPM) and *On-the-fly Gradient Modulation* (OGM). These two strategies respectively target the feed-forward and back-propagation stages. Specifically, by monitoring the discriminative discrepancy between modalities during training, OPM drops the feature of dominant modality with dynamical probability while OGM on-the-fly mitigates its gradient, thereby improving the learning of the less discriminative modality. As shown in Fig 1, the previously worse learnt visual modality of the VGGSound dataset shows marked improvement after applying either OPM or OGM (blue and green lines in Fig 1(a) and Fig 1(b)). The performance of the dominant audio modality also benefits. Furthermore, our methods notably enhance the overall multimodal model with the joint training strategy (see Fig 1(c)). To thoroughly demonstrate the effectiveness and versatility of OPM and OGM, we evaluate their performance across various multimodal tasks, achieving consistent improvement. Moreover, our modulation methods help improve multimodal representation by enhancing the learning of each modality.

Our previous conference paper [15] has exposed the imbalanced learning phenomenon in the back-propagation stage and achieved considerable performance with the proposed on-the-fly gradient modulation method. In this paper, we provide a more systematic analysis of the imbalanced learning problem from both the feed-forward and the back-propagation stages, and further introduce the on-the-fly prediction modulation method, which focuses on the feed-forward stage. These two strategies are designed to holistically consider the stages of optimization. Moreover, we additionally provide an analysis that how our improvement in representation contributes to better model performance. In addition, we further expand our methods to address more complex cross-modal interactions and a broader range of multimodal tasks. In experiments, we conduct extensive evaluation across a diverse set of modalities, encompassing various numbers and types. A wide range of fine-grained analysis and ablation studies are also conducted to validate our methods thoroughly. Overall, our main contributions are as follows:

- We observe and analyze the imbalanced learning phenomenon that the uni-modal encoders in multimodal model are imbalanced under-optimized and one modality could be worse learnt than others during training, from both the feed-forward and the back-propagation stages.
- OPM and OGM methods are proposed to ease imbalanced learning problem by controlling the optimization of each modality adaptively in the feed-forward and the back-propagation stages, respectively.
- OPM and OGM can be equipped with various multimodal tasks and models, demonstrating their promising effectiveness and versatility in diverse multimodal learning scenarios.

2 Related Works

2.1 Multimodal learning

Multimodal learning, which integrates information from multiple modalities, has been attracting increasing attention due to the growing amount of multimodal data. This data naturally contains correlated information from diverse sources. Recently, the field of multimodal learning has witnessed rapid development. On the one hand, multimodal modalities are used to enhance the performance of existing uni-modal tasks, such as multimodal action recognition [3], [4], [16]

and audio-visual speech recognition [17], [18]. On the other hand, more researchers also begin to explore and solve new multimodal problems and challenges, like audio-visual event localization [7], [19] and multimodal question answering [20]. To efficiently learn and integrate multiple modalities, most multimodal methods tend to use the joint training strategy, which optimizes different modalities with a uniform learning objective. However, this approach may not fully exploit all modalities, causing some to be less effectively learned than others. This can prevent multimodal models from achieving their expected performance, even though they are superior to their uni-modal counterparts. In this paper, we propose on-the-fly modulation methods to improve the joint learning of multimodal models via dynamically controlling the uni-modal optimization.

2.2 Imbalanced multimodal learning

The multimodal model is expected to outperform its uni-modal counterpart since it takes data containing information from multiple views. But the widely used joint training multimodal model does not always work well based on existing studies [8], which prompts researchers to investigate the reasons. Recent studies point out that the jointly trained multimodal model cannot effectively improve the performance with more information as expected due to the discrepancy between modalities [8], [9], [10], [11], [21]. Wang et al. [8] found that multiple modalities often converge and generalize at different rates, thus training them jointly with a uniform learning objective is sub-optimal, leading to the multimodal model sometimes is inferior to the uni-modal ones. Also, Winterbottom et al. [21] indicated an inherent bias in the TVQA dataset towards the textual subtitle modality. Besides the empirical observation, Huang et al. [10] further theoretically proved that the jointly trained multimodal model cannot efficiently learn features of all modalities, and only a subset of them can capture sufficient representation. They called this process "Modality Competition". In the recent past, several methods have emerged attempting to alleviate this problem [8], [11], [22], [23]. Wang et al. [8] proposed to add additional uni-modal loss functions besides the original multimodal objective to balance the training of each modality. Du et al. [22] utilized the well-trained uni-modal encoders to improve the multimodal model by knowledge distillation. Wu et al. [11] measured the speed at which the model learns from one modality relative to the other modalities, and then proposed to guide the model to learn from previously underutilized modalities. Wei et al. [24] introduced a Shapley-based sample-level modality valuation metric, to observe and alleviate the fine-grained modality discrepancy. Wei et al. [25] further considered the possible limited capacity of modality and utilized the re-initialization strategy to control uni-modal learning. Differently, Yang et al. [26] focused on the influence of imbalanced multimodal learning on multimodal robustness, and proposed a robustness enhancement strategy. While these methods have improved multimodal learning, a comprehensive analysis of imbalanced multimodal learning is still lacking. In this paper, we begin with a systematic analysis of both the feed-forward and back-propagation stages to understand how multimodal discrepancies impact training. Based on the analysis, we propose to alleviate these issues by adaptively controlling the optimization of each modality without introducing additional modules.

2.3 Modality dropout

In our OPM method, we adaptively drop the feature of the dominant modality during training to enhance the learning of the remaining modalities. Following the regularization technique dropout [27], different network dropout strategies are proposed and show their effectiveness [28], [29], [30]. In recent years, the idea of dropout has been transferred into the multimodal learning area to drop modalities during training [31], [32], [33], [34], [35]. Neverova et al. [31] proposed the ModDrop method that drops modalities with a certain probability during training to break the dependency between modalities and improve model robustness to missing modalities. Xiao et al. [35] claimed that dropping the modality with a faster learning pace could slow down its converging and facilitate the training of multimodal network. However, the previous modality dropout methods usually fix the drop probability of each modality during the whole training process [31], [35], which can not well match the dynamic learning process of multiple modalities, especially in the case that modalities vary in learning pace. Hence, our OPM method adaptively adjusts the drop probability of each modality during training by monitoring the discriminative discrepancies between modalities, thereby focusing more on the less discriminative modalities.

2.4 Generalization and stochastic gradient noise

Based on the existing studies, the gradient noise in SGD is considered to have an essential correlation with the generalization ability of deep models [36], [37], [38], [39], [40]. This stochastic gradient noise brought by random mini-batch sampling, is believed that can serve as regularization and assist the deep model to escape from saddle point or local optimum [37], [38], [39], [41]. Jin et al. [42] proposed that the suitable perturbation or noise in the gradient can help the model to escape saddle points efficiently. Neelakantan et al. [43] demonstrated that adding noise to gradient is helpful to potentially improve the training of deep neural networks. Zhou et al. [44] further provided theoretical proof that the stochastic gradient algorithms with proper Gaussian noise, are guaranteed to converge to the global optimum in polynomial time with random initialization. In our OGM method, to enhance the generalization ability of the multimodal model, we introduce extra Gaussian noise into the modified gradient.

3 Method and Analysis

3.1 Imbalanced learning analysis

As demonstrated in Fig 1, the uni-modal encoders in the jointly trained multimodal model are under-optimized to a different degree, and some modalities are worse learnt than others. In this section, we analyze this imbalanced phenomenon and find that the modality with more discriminative information dominates the optimization progress of multimodal model, causing other modalities to be worse

under-optimized. In the analysis, we consider the widely-used late-fusion multimodal model. It should be noted that our proposed methods are also applicable to more complex cross-modal interactions in practice (as demonstrated by experiments in Sec 4.3), although our analysis focuses on the general late-fusion multimodal model. Each modality is processed by the corresponding uni-modal encoder. Then their features are fused by the concatenation operation and passed to a single-layer linear classifier to produce the final prediction. The cross-entropy function is used as the discriminative learning objective.

For convenience, the dataset is denoted by $S = \{(x_i, y_i)\}_{i=1,2...N}$. Suppose the number of modalities is M. Each x_i contains inputs of M modalities: $x_i = (x_i^1, x_i^2, \cdots, x_i^M)$. $y_i \in \{1, 2, \cdots, C\}$ is the target label of sample x_i and C is the number of categories. For modality m, where $m \in \{1, 2, \cdots, M\}$, its input is processed by the corresponding encoder $\varphi^m(\theta^m, \cdot)$. θ^m are the parameters of encoder. After extraction, their features are fused via concatenation, and passed to a single-layer linear classifier. $W \in \mathbb{R}^{C \times \sum_{m=1}^{M} d_{\varphi^m}}$ and $b \in \mathbb{R}^C$ denote the parameters of the linear classifier. d_{φ^m} is the output dimension of $\varphi^m(\theta^m, \cdot)$.

3.1.1 Feed-forward stage

With the above notions, the logits output of the considered multimodal model can be formulated as follows:

$$f(x_i) = W[\varphi^1(\theta^1, x_i^1); \varphi^2(\theta^2, x_i^2); \cdots; \varphi^M(\theta^M, x_i^M)] + b. \tag{1}$$

To observe the uni-modal components individually, we can mathematically transform the calculation of output $f(x_i) \in \mathbb{R}^C$, and then Eqn 1 is rewritten as:

$$f(x_i) = W^1 \cdot \varphi_i^1 + W^2 \cdot \varphi_i^2 + \cdots + W^M \cdot \varphi_i^M + b, \tag{2}$$

where $\varphi^m(\theta^m, x_i^m)$ is denoted as φ_i^m for simplicity. W is divided into M blocks: $[W^1; W^2; \cdots; W^M]$. $W^m \in \mathbb{R}^{C \times d_{\varphi^m}}$. Based on Eqn 2, the prediction of multimodal model is determined by *the sum of uni-modal components*, $W^m \cdot \varphi_i^m$, in the feed-forward stage.

3.1.2 Back-propagation stage

Here we consider the cross-entropy loss function and the *Gradient Descent* (GD) optimization method. The loss of sample x_i is $\ell(x_i, y_i) = -\log \frac{e^{f(x_i)_{y_i}}}{\sum_{c=1}^{C} e^{f(x_i)_c}}$, where C is the number of categories, and $f(x_i)_c$ is the logits for class c. During optimization, for modality m, where $m \in \{1, 2, \cdots, M\}$, W^m and $\varphi^m(\theta^m, \cdot)$ are updated as:

$$\begin{aligned} W_{t+1}^m &= W_t^m - \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_{W^m} \ell(x_i, y_i) \\ &= W_t^m - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell(x_i, y_i)}{\partial f(x_i)} \varphi_{i,t}^m, \end{aligned} \tag{3}$$

$$\begin{aligned} \theta_{t+1}^m &= \theta_t^m - \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta^m} \ell(x_i, y_i) \\ &= \theta_t^m - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell(x_i, y_i)}{\partial f(x_i)} \frac{\partial (W_t^m \cdot \varphi_{i,t}^m)}{\partial \theta_t^m}, \end{aligned} \tag{4}$$

where η is the learning rate. Referring to Eqn 3 and Eqn 4, the update of W^m and parameters in φ^m has no correlation with the other modality, except the term related to the loss, *i.e.*, $\frac{\partial \ell(x_i, y_i)}{\partial f(x_i)}$. The uni-modal encoders thus are hard to make adjustments according to the feedback from each other. Denote logit of class c is denoted as $f(x_i)_c$. Then, the gradient $\frac{\partial \ell(x_i, y_i)}{\partial f(x_i)_c}$ for category c can be written as:

$$\frac{\partial \ell(x_i, y_i)}{\partial f(x_i)_c} = \frac{e^{(W^1 \cdot \varphi_i^1 + W^2 \cdot \varphi_i^2 + \cdots + W^M \cdot \varphi_i^M + b)_c}}{\sum_{j=1}^{C} e^{(W^1 \cdot \varphi_i^1 + W^2 \cdot \varphi_i^2 + \cdots + W^M \cdot \varphi_i^M + b)_j}} - 1_{c=y_i}, \tag{5}$$

According to Eqn 5, the term $1_{c=y_i}$ is constant and not related to specific modality. For the first term, for each category, its denominator is the same. And the value of its molecule is determined by the sum of uni-modal components, $W^m \cdot \varphi_i^m$, for category c. Therefore, the concrete value of gradient for category c, $\frac{\partial \ell(x_i, y_i)}{\partial f(x_i)_c}$, is also controlled by *the sum of uni-modal components*, although it is not analytically equal to the sum of corresponding uni-modal components.

3.1.3 Dominated optimization process

The recent study has empirically shown that the different modalities could vary in the optimization process [8]. Meanwhile, our analysis of the feed-forward and back-propagation stages shows that both the multimodal prediction and the gradient values are controlled by *the sum of uni-modal components*. Therefore, when one modality, such as modality m, has more discriminative information, it would dominate the multimodal prediction $f(x_i)$ and gradient $\frac{\partial \ell(x_i, y_i)}{\partial f(x_i)}$ via $W^m \cdot \varphi_i^m$. Even if another modality is under-optimized and yields incorrect results, the component from the better-performing modality m can still "correct" these errors during summation, thus influencing both the feed-forward and back-propagation stages. Therefore, with Eqn 2 and Eqn 5, another modality still with relatively lower confidence about the correct category, only earns limited optimization efforts, leading to it being underutilized. Overall, based on the above analysis, the modality with better performance dominates the optimization progress. Inevitably, as the multimodal model approaches convergence, the less discriminative modalities could still require further training due to their under-optimized features.

Additionally, the analysis in this section is based on the multimodal model that uses a single-layer classifier. For more general cases, we extend this analysis to the multimodal model that uses a multi-layer classifier with non-linear activation function in Appendix A.

3.2 On-the-fly modulation strategies

3.2.1 On-the-fly prediction modulation

In the multimodal dataset, there often exists a dominant modality with more discriminative information. The overall performance of the model tends to be more dependent on the modality with more discriminative information, which in turn influences the optimization of other modalities. Hence, to weaken the reliance on the dominant modality, we propose randomly dropping the feature of the more discriminative modality with a specific probability during the feed-forward stage, thus specifically accelerating the training

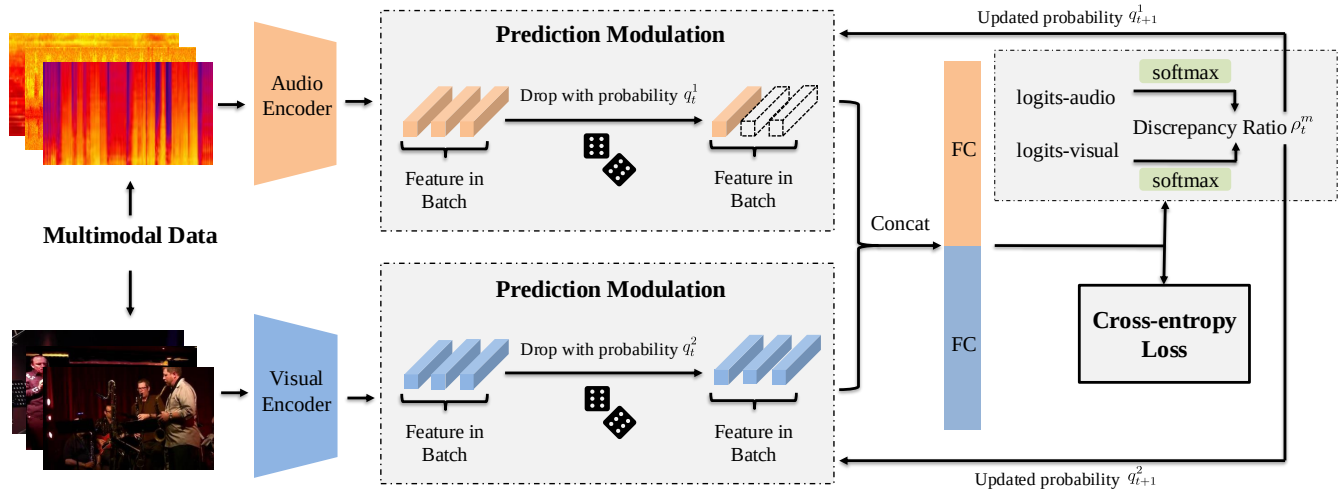


Fig. 2. **The pipeline of the On-the-fly Prediction Modulation.** Here we take two modalities as examples. In the feed-forward stage, the feature of modality m is randomly dropped with probability q^m, where the probability is determined by the discriminative discrepancy ratio at the last iteration. Via OPM, the remained feature of suppressed modality could affect the multimodal prediction more, accordingly improving its learning.

Algorithm 1 Multimodal learning with OPM strategy

Input: Training dataset $S = \{(x_i, y_i)\}_{i=1,2...N}$, iteration number T, initialized modal-specific parameters θ^m, $m \in \{1, 2, \cdots, M\}$.
for $t = 0, \cdots, T-1$ **do**
  Sample a fresh mini-batch B_t from S;
  Feed-forward the batched data B_t to the model;
  Calculate ρ_t^m using Eqn 6 and Eqn 7;
  Calculate q_{t+1}^m using Eqn 8;
  Drop the input of each modality with probability q_t^m;
  Calculate gradient using back-propagation;
  Update the model parameters.
end for

of the suppressed modality. What's more, the discriminative ability of uni-modal features is gradually improved but at a different rate during training. Hence, the discrepancy in the discriminative ability between uni-modal features is dynamic. Correspondingly, it is necessary to adaptively adjust the drop probability of each modality during the training. Overall, in the proposed OPM method, the drop probability of modality with more discriminative information is adaptively adjusted during training based on the discriminative discrepancy degree between modalities. The pipeline of our OPM method is shown in Fig 2.

Here we follow the notation in Sec 3.1. To monitor the discriminative discrepancy between modalities during training, we first propose to estimate the uni-modal discriminative performance via:

$$s_i^m = \sum_{c=1}^{C} 1_{c=y_i} \cdot \iota(W_i^m \cdot \varphi_i^m(\theta^m, x_i^m) + \frac{b}{M})_c, \tag{6}$$

where ι is the softmax function and y_i is the ground truth label of sample x_i. As stated in Eqn 6, for modality m, the uni-modal component in the final multimodal prediction, $(W_i^m \cdot \varphi_i^m(\theta^m, x_i^m) + \frac{b}{M})$, is used as its the approximated prediction. Here we split the bias term into $\frac{b}{M}$ to estimate uni-modal performance. Since the bias term often has less effect on the prediction, its split could not have a great influence on the estimation. In Sec 4.5.2, we provided ablation studies about the split of bias term.

Since the modality with more discriminative information tends to have higher confidence for the correct category, *i.e.*, the value of s tends to be higher, then we propose to measure the discriminative discrepancy ratio of modality m to other modalities by:

$$\rho_t^m = \frac{1}{M-1} \sum_{j \in [M], j \neq m} \frac{\sum_{i \in B_t} s_i^m}{\sum_{i \in B_t} s_i^j}. \tag{7}$$

When the average uni-modal discriminative performance ratio to other modalities is larger than 1, *i.e.*, $\rho_t^m > 1$, modality m is more discriminative. B_t is a random mini-batch which is chosen in the t-th step.

With ρ_t^m to dynamically monitor the discriminative discrepancy among modalities, we can adaptively adjust the drop probability of modality m through:

$$q_{t+1}^m = \begin{cases} q_{base} \cdot (1 + \lambda \cdot z(\rho_t^m)) & \rho_t^m > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where the base drop probability is q_{base} and $z(\cdot)$ is a monotonically increasing function with a value range between 0 and 1. Hyper-parameter q_{base} ranging in $(0, 1)$ controls the range of modality dropout probability and $\lambda > 0$ determines the degree of adjustment. Based on the modulation of OPM, for modality with more discriminative information ($\rho_t^m > 1$), the discrepancy degree, *i.e.*, ρ_t^m, is processed by $z(\cdot)$ to map it into $(0, 1)$ as the increase of base drop probability. The drop probability of less discriminative one ($\rho_t^m \leq 1$) is set to 0.

When $\rho_t^m > 1$, the concrete value of drop probability q_{t+1}^m ranges in $(q_{base}, q_{base} \cdot (1 + \lambda))$. In experiments, the maximum value of q_{t+1}^m is set to 1 to avoid illegal value. $tanh(x - 1)$ function is used as $z(x)$[2]. In addition, when

2. More alternative strategies and analysis are provided in Sec 4.5.1.

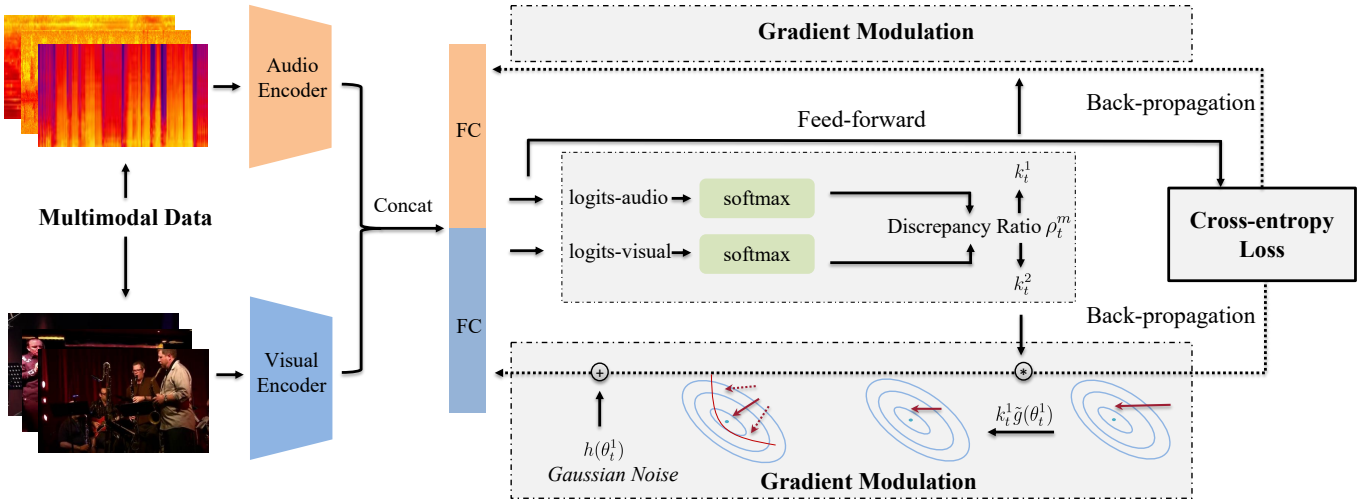


Fig. 3. **The pipeline of the On-the-fly Gradient Modulation strategy.** Here we take two modalities as example. In the back-propagation stage, the gradient of modality m is modulated with k^m, which is determined by the discriminative discrepancy ratio at this iteration. Via OGM, the gradient of modality with more discriminative information is weakened, while the remained modality is not affected and can gain more training.

Algorithm 2 Multimodal learning with OGM strategy

Input: Training dataset $S = \{(x_i, y_i)\}_{i=1,2...N}$, iteration number T, initialized modal-specific parameters θ^m, $m \in \{1, 2, \cdots, M\}$.

for $t = 0, \cdots, T-1$ **do**

Sample a fresh mini-batch B_t from S;

Feed-forward the batched data B_t to the model;

Calculate ρ_t^m using Eqn 6 and Eqn 7;

Calculate k_t^m using Eqn 11;

Calculate gradient $\tilde{g}(\theta_t^m)$ using back-propagation;

Sample $h(\theta_t^m)$ based on covariance of gradient $\tilde{g}(\theta_t^m)$;

Update using $\theta_{t+1}^m = \theta_t^m - \eta(k_t^m \tilde{g}(\theta_t^m) + h(\theta_t^m))$.

end for

the discrepancy between modalities is tiny, *i.e.*, ρ_t^m is close to 1, the dropout probability q_{t+1}^m is close to q_{base}, since $tanh(\rho_t^m - 1)$ is approaching 0.

The overall OPM method is provided in Alg 1. The OPM method alleviates the imbalanced learning problem in the feed-forward stage. Via the proposed strategy, the effect on multimodal prediction of modality with better performance is relatively weakened. The larger the discriminative discrepancy, the stronger the modulation. Consequently, the previously suppressed modality could have a greater influence on multimodal prediction, thereby improving its learning. In addition, based on recent study [45], the dropout strategy introduces additional noise into the gradient, which can improve model generalization. This is because gradient noise can help the model converge to wider minima, which typically generalize better. Therefore, the OPM method could also bring better multimodal generalization ability.

3.2.2 On-the-fly gradient modulation

The OPM method enhances the optimization of suppressed modality in the feed-forward stage. As discussed in Sec 3.1, the modality with more discriminative information also dominates the gradient during the back-propagation stage, leading to lower loss, and then limiting the gradient of other modalities. Then, the OGM modulation strategy is proposed to amend the optimization of each modality in the back-propagation stage by mitigating the gradient of more discriminative modality based on the modality discrepancy during training. The pipeline of OGM is shown in Fig 3.

Specifically, when using GD optimization method, for modality m, the parameters θ^m of the encoder φ^m is updated as follows:

$$\theta_{t+1}^m = \theta_t^m - \eta \nabla_{\theta^m} \mathcal{L}(\theta_t^m), \tag{9}$$

where $\nabla_{\theta^m} \mathcal{L}(\theta_t^m) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta^m} \ell(x_i; \theta_t^m)$ is the full gradient over all training samples. ℓ is the loss function.

In practice, we use the widely used *Stochastic Gradient Descent* (SGD) optimization method, the parameters are updated as:

$$\theta_{t+1}^m = \theta_t^m - \eta \tilde{g}(\theta_t^m), \tag{10}$$

where $\tilde{g}(\theta_t^m) = \frac{1}{|B_t|} \sum_{x \in B_t} \nabla_{\theta^m} \ell(x; \theta_t^m)$ is the gradient of current mini-batch B_t, and it can be considered as an unbiased estimation of the full gradient $\nabla_{\theta^m} \mathcal{L}(\theta_t^m)$ [41]. $|B_t|$ is the number of samples in this mini-batch.

Specific to the imbalanced learning problem, OGM proposes to control the uni-modal optimization in the back-propagation stage by modulating the gradient of each modality. Concretely, the gradient of modality with more discriminative information is mitigated, then the overall training process is slowed down and another modality could gain more optimization efforts. Similarly, considering the discriminate discrepancy is dynamic in the training process, the degree of gradient mitigation is adaptively adjusted. As in Sec 3.2.1, the discriminative discrepancy is also formulated as ρ_t^m in Eqn 7. By means of ρ_t^m, we can dynamically modulate the gradient by:

$$k_t^m = \begin{cases} 1 - \alpha \cdot z(\rho_t^m) & \rho_t^m > 1 \\ 1 & \text{otherwise,} \end{cases} \tag{11}$$

where $\alpha > 0$ is a hyper-parameter to control the degree of gradient modulation and $z(\cdot)$ is a monotonically increasing

function with a value range between 0 and 1. Then, when $\rho_t^m > 1$, the concrete value of k_t^m ranges in $(1-\alpha, 1)$. In experiments, the minimum value of k_t^m is set to 0 to avoid illegal value. Also, $tanh(x-1)$ function is used as $z(x)$ in experiments. After integrating the coefficient k_t^m, θ_t^m in iteration t is updated as follows:

$$\theta_{t+1}^m = \theta_t^m - \eta \cdot k_t^m \tilde{g}(\theta_t^m). \tag{12}$$

After modulation, the gradient of modality with better performance ($\rho_t^m > 1$) is mitigated, and the other modality is not affected. The larger the discriminative discrepancy, the stronger the modulation. Correspondingly, the overall training process is slowed down and the suppressed modality can gain more training, easing the imbalanced learning problem. In addition, when the discrepancy between modalities is tiny, *i.e.*, ρ_t^m is close to 1, the gradient modulation coefficient k_t^m is also close to 1, since $tanh(\rho_t^m - 1)$ is approaching 0. In this case, the gradient of relatively better modality is only slightly mitigated.

However, since the gradient mitigation operation could potentially harm the model generalization ability as the following analysis [41], we further introduce additional noise into the gradient to enhance generalization.

Concretely, as stated before, gradient of mini-batch B_t, $\tilde{g}(\theta_t^m)$, is an un-biased estimation of full gradient $\nabla_{\theta^m}\mathcal{L}(\theta_t^m)$. For a sufficiently large batch size, based on the central limit theorem, the gradient of each mini-batch is approximately Gaussian [41]:

$$\tilde{g}(\theta_t^m) \sim \mathcal{N}(\nabla_{\theta^m}\mathcal{L}(\theta_t^m), \frac{U}{|B_t|}), \tag{13}$$

where U is the covariance matrix, brought by the random sampling of SGD. Then the parameter update of SGD can be rewritten as follows:

$$\begin{aligned} \theta_{t+1}^m &= \theta_t^m - \eta\tilde{g}(\theta_t^m), \\ \theta_{t+1}^m &= \theta_t^m - \eta\nabla_{\theta^m}\mathcal{L}(\theta_t^m) + \xi_t, \xi_t \sim \mathcal{N}(0, \frac{\eta^2}{|B_t|}U), \end{aligned} \tag{14}$$

where ξ_t is considered as the SGD noise term. Based on the existing studies [36], [41], this noise term in SGD has a close relationship with model generalization ability. The larger SGD noise often tends to bring better generalization, since it could help the model converge at a wider minima. Based on Eqn 14, we can have that the strength of SGD noise term ξ_t is proportional to the ratio of learning rate to batch size. To verify that larger noise has better generalization, we conduct experiments with different $|B_t|$ and η. As shown in Tab 1, larger η or smaller $|B_t|$ indeed bring better multimodal model performance.

With the above analysis, in OGM method, when modulating the gradient via coefficient k_t^m, θ^m is updated as:

$$\begin{aligned} \theta_{t+1}^m &= \theta_t^m - \eta \cdot k_t^m \tilde{g}(\theta_t^m), \\ \theta_{t+1}^m &= \theta_t^m - \eta \cdot k_t^m \nabla_{\theta^m}\mathcal{L}(\theta_t^m) + \xi_t', \\ \xi_t' &\sim \mathcal{N}(0, (k_t^m)^2 \cdot \frac{\eta^2}{|B_t|}U). \end{aligned} \tag{15}$$

According to Eqn 11, when modality m is more discriminative with $\rho_t^m > 1$, the modulation coefficient $k_t^m < 1$. Inevitably, when the learning rate and batch size are fixed, the modulated SGD noise term ξ_t' is smaller than the original ξ_t, leading to the generalization ability of multimodal model could be affected. Therefore, it is desirable to recover the generalization ability.

TABLE 1
Experiments about different batch size $|B_t|$ and learning rate η on CREMA-D and Kinetics-Sounds dataset.

Method	CREMA-D		Kinetics-Sounds	
	Acc	mAP	Acc	mAP
$\|B_t\| = 32, \eta = 1e-5$	40.5	48.5	47.6	51.4
$\|B_t\| = 32, \eta = 1e-4$	63.4	69.1	61.9	65.7
$\|B_t\| = 32, \eta = 1e-3$	**66.9**	**72.1**	**63.0**	**67.3**
$\|B_t\| = 64, \eta = 1e-3$	64.7	71.4	62.1	66.4
$\|B_t\| = 128, \eta = 1e-3$	61.9	66.9	58.7	63.2

To avoid this potential generalization reduction issue, we propose to introduce a randomly sampled Gaussian noise $h(\theta_t^m) \sim \mathcal{N}(0, \frac{U}{|B_t|})$ into the gradient to recover the strength of noise term:

$$\begin{aligned} \theta_{t+1}^m &= \theta_t^m - \eta \cdot (k_t^m \tilde{g}(\theta_t^m) + h(\theta_t^m)) \\ &= \theta_t^m - \eta \cdot k_t^m \nabla_{\theta^m}\mathcal{L}(\theta_t^m) + \xi_t' + \epsilon_t. \end{aligned} \tag{16}$$

The additional Gaussian noise $h(\theta_t^m)$ has a zero mean and the same covariance as current $\tilde{g}(\theta_t^m)$ at iteration t. Then, $\epsilon_t \sim \mathcal{N}(0, \frac{\eta^2}{|B_t|}U)$. Since ϵ_t and ξ_t' are two independent variants, and can be combined as a single term, then Eqn 16 can be rewritten as:

$$\begin{aligned} \theta_{t+1}^m &= \theta_t^m - \eta \cdot k_t^m \nabla_{\theta^m}\mathcal{L}(\theta_t^m) + \xi_t'', \\ \xi_t'' &\sim \mathcal{N}(0, ((k_t^m)^2 + 1) \cdot \frac{\eta^2}{|B_t|}U). \end{aligned} \tag{17}$$

Hence, the SGD noise term ξ_t'' is recovered and even enhanced, compared with the original ξ_t, avoiding the risk of harming multimodal model generalization. The overall OGM method is provided in Alg 2. The OGM method eases the imbalanced learning problem in the back-propagation stage. Using the proposed strategy, the gradient of modality with better performance is weakened with guaranteed generalization ability based on the discriminative discrepancy, providing the suppressed modality with more training.

3.2.3 Model performance analysis

Using our on-the-fly modulation methods, we ease the imbalanced multimodal learning problem by enhancing uni-modal learning. Then, our methods improve the quality of uni-modal features. Since the multimodal representation is the fusion of uni-modal features, the improvement of uni-modal feature quality is expected to bring the improvement of multimodal latent representation. In Appendix B, we provide an analysis of how this improvement in representation attribute to better multimodal performance. Both quantitative and qualitative analysis are provided to verify how our methods improve the multimodal representation.

4 Experiment

4.1 Dataset

CREMA-D [47] is an emotion recognition dataset with two modalities: audio and visual. This dataset contains 7,442 video clips of 2-3 seconds for 91 persons speaking short works with different emotions. It covers 6 most common

TABLE 2
Combined with different fusion methods. Encoders of UCF-101 dataset are pre-trained on ImageNet. OF denotes Optical Flow modality. Audio/RGB-only and Visual/OF-only methods are individually trained uni-modal model.

Method	CREMA-D (Audio+Visual)		Kinetics-Sounds (Audio+Visual)		UCF-101 (RGB+Optical Flow)		VGGSound (Audio+Visual)	
	Acc	mAP	Acc	mAP	Acc	mAP	Acc	mAP
Audio/RGB-only	61.4	67.7	49.9	51.3	77.2	82.5	47.5	49.3
Visual/OF-only	50.9	52.2	46.7	48.1	59.9	64.1	28.7	28.3
Concatenation	66.9	72.1	63.0	67.3	80.5	86.4	52.7	54.9
Concatenation-OPM	**75.1**	81.2	**67.0**	**72.5**	**81.9**	**88.0**	**54.1**	**56.5**
Concatenation-OGM	74.6	**81.4**	65.4	71.4	81.5	87.6	53.6	55.9
Summation	60.8	65.8	62.7	66.0	80.9	86.8	52.5	54.7
Summation-OPM	**63.4**	**70.7**	**66.5**	**72.3**	**81.9**	**88.0**	**53.8**	**56.2**
Summation-OGM	**63.4**	69.8	65.4	70.8	81.3	87.6	52.9	54.9
FiLM [46]	61.8	66.6	60.6	64.4	75.2	81.0	49.8	51.6
FiLM-OPM	62.4	**69.4**	64.5	68.8	**75.7**	81.3	**51.3**	**53.5**
FiLM-OGM	**63.0**	68.3	**65.3**	**70.9**	**75.7**	**81.6**	50.0	51.3

emotions. In the experiments, all samples are randomly divided into a 6,698-sample training and validation set and a 744-sample testing set.

Kinetics-Sounds [48] is an action recognition dataset with two modalities, audio and visual. It contains 31 human action classes selected from Kinetics dataset [49]. All videos are manually annotated utilizing Mechanical Turk and cropped to 10 seconds long around the action. This dataset contains 19k 10-second video clips. In our experiments, we follow the original dataset division.

UCF-101 [50] is an action recognition dataset with two modalities, RGB and optical flow. It has 13,320 videos from 101 action categories. UCF-101 dataset contains two modalities: RGB and optical flow. The entire dataset is divided into a 9,537-sample training set and a 3,783-sample test set according to the original setting. **UCF-101-Three** dataset introduces the additional RGB-Difference modality based on the UCF-101 dataset

VGGSound [13] is an event recognition dataset with two modalities, audio and visual. It contains over 200k clips for 309 different classes, covering a wide range of daily audio events. Each video has a duration of 10 seconds. In our experiment, the dataset division is the same as [13]. 168,618 videos are used for training and validation, and 13,954 videos are used for testing.

CMU-MOSI [51] is a sentiment analysis dataset with three modalities, audio, video and text. It is annotated with utterance-level sentiment labels. In experiments, labels are binary to classify whether the sentiment is positive or negative. This dataset consists of 93 movie review videos segmented into 2,199 utterances. The division of dataset follows the official split.

AVE [7] is a video dataset for the audio-visual event localization task. It contains 4,143 10-second videos from 28 event categories. Each video consists of at least one 2-second long audio-visual event, covering a wide range of domains, including human activities, animal activities, music performances, and vehicle sounds. In our experiments, the division of the dataset follows the official split.

MUSIC-AVQA [52] is designed for audio-visual question answering under musical scenario. It contains 9,288 videos covering 22 instruments, with a total duration of over 150 hours and 45,867 Question-Answering pairs. Each video contains around 5 QA pairs on average. In experiments, we follow the official split for training, evaluation, and test sets.

4.2 Experimental settings

In our experiments, when not specified, ResNet-18 [53] is used as the backbone. Concretely, for the visual encoder, we take multiple frames as the input, and feed them into the 2D network like [54] does; for the audio encoder, we modified the input channel of ResNet-18 from three to one like [13] does and the rest parts remain unchanged; for the optical flow encoder, we stack the horizontal vector and vertical vector as one frame, then multiple frames are also put into the 2D network as [54] does. In addition, the encoders used for UCF-101 and UCF-101-Three are ImageNet pre-trained and encoders of other datasets are trained from scratch. For the CMU-MOSI dataset, transformer-based networks are used as the backbone and trained from scratch.

Videos in Kinetics-Sounds, UCF-101, UCF-101-Three and VGGSound datasets are extracted at 1fps and three frames are uniformly sampled as the visual input. Videos in the AVE dataset are extracted frames with 1fps and all ten frames are used as the visual input. The audio data is first re-sampled into 16KHz and transformed into a spectrogram with size $257 \times 1,004$ using a window with length of 512 and overlap of 353. For the CREMA-D dataset with only 2-3 seconds video, One visual frame is randomly extracted from each clip and the audio data is processed into a spectrogram of size 257×299 with a window length of 512 and overlap of 353.

During training, we use SGD with momentum (0.9) as the optimizer. For CREMA-D, Kinetics-Sounds, UCF-101, UCF-101-Three, and VGGSound datasets, the learning rate is $1e-3$, and the batch size is 32. For CMU-MOSI, AVE, and MUSIC-AVQA datasets, the learning rate is $1e-4$, and batch size is 64. All models are trained on 2 NVIDIA RTX 3090 (Ti).

Evaluation metric. For multi-class classification tasks, the widely-used **Accuracy** and **mAP** are used as evaluation metric:

$$\text{Acc} = \frac{\sum_{i=1}^{A} 1_{\hat{y_i}=y_i}}{A}, \tag{18}$$

where A is the number of testing samples, and $\hat{y_i}$ is the

TABLE 3
Combined with cross-modal interaction methods on the Kinetics-Sounds and UCF-101 dataset.

Method	Kinetics-Sounds Acc	Kinetics-Sounds mAP	UCF-101 Acc	UCF-101 mAP
CentralNet [12]	66.4	70.6	82.0	87.4
CentralNet-OPM	68.3	73.5	82.9	88.1
CentralNet-OGM	**68.8**	**73.9**	**83.2**	**88.8**
VATT [55]	61.8	64.0	81.2	86.5
VATT-OPM	**69.2**	**76.0**	**82.6**	**89.2**
VATT-OGM	67.6	73.3	82.3	88.0
MMTM [56]	63.8	68.4	79.8	85.3
MMTM-OPM	**67.9**	**73.8**	**80.5**	**86.7**
MMTM-OGM	66.1	71.5	79.9	85.8

prediction of model for sample x_i in testing set.

$$\begin{aligned} \mathrm{AP}_c &= \sum_n (R_{c,n} - R_{c,n-1}) P_{c,n}, \\ \mathrm{mAP} &= \frac{1}{C} \sum_{c=1}^{C} \mathrm{AP}_c, \end{aligned} \tag{19}$$

where C is the category number. For class c, AP_c summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. $P_{c,n}$ and $R_{c,n}$ are the precision and recall for class c at the $n-$th threshold. For the binary classification task, sentiment analysis, **Accuracy** and **F1 score** are used as evaluation metric:

$$\text{F1 score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{20}$$

In practice, the widely-used Python package scikit-learn is used for evaluation.

4.3 Comparison on the multimodal task

4.3.1 Combination with different fusion methods

We first apply OPM and OGM methods to two vanilla fusion methods: Concatenation and Summation. Additionally, we compared the specifically designed fusion method FiLM [46]. The results on four datasets are as shown in Tab 2. We also provide the performance of individually trained uni-modal models for comparison. It can be observed that the uni-modal performance is imbalanced across different datasets, demonstrating that the discriminative ability of different modalities varies on different datasets. For instance, the performance of audio-only model outperforms the visual-only model on the sound-oriented VGGSound dataset. Furthermore, in some cases, the jointly trained multimodal model can be inferior to the best performing uni-modal models. For example, the performance of the audio-only model on CREMA-D dataset is better than the multimodal model with Summation fusion.

Moreover, OPM and OGM can bring considerable improvement to both vanilla and specifically designed fusion methods, which demonstrates the effectiveness and satisfactory flexibility of our strategies. As shown in Fig 4 and Tab 4, although our modulation for dominant modality slows down the overall training process a bit, our OPM and OGM methods do not bring much additional training time with the same training epoch, compared with Concatenation baseline. Additionally, it can be observed that the convergence loss of our OPM and OGM methods is higher than that of

TABLE 5
Comparison with other modulation methods on CREMA-D, Kinetics-Sounds, and UCF-101 dataset. OPM and OGM methods are based on Concatenation fusion.

Method	CREMA-D Acc	CREMA-D mAP	Kinetics Sounds Acc	Kinetics Sounds mAP	UCF-101 Acc	UCF-101 mAP
Concatenation	66.9	72.1	63.0	67.3	80.5	86.4
GBlending [8]	72.0	78.6	**67.5**	72.1	80.9	86.6
Greedy [11]	68.4	76.8	65.2	69.6	80.9	86.9
OPM	**75.1**	81.2	67.0	**72.5**	**81.9**	**88.0**
OGM	74.6	**81.4**	65.4	71.4	81.5	87.6

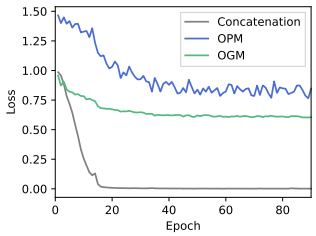


Fig. 4. Training loss change on CREMA-D dataset.

TABLE 4
Training time on CREMA-D dataset.

Method	Training time (Min)
Concatenation	1.00×
GBlending [8]	1.50×
Greedy [11]	1.01×
OPM	1.01×
OGM	1.01×

the baseline, possibly because our modulation prevents the multimodal model from over-memorizing the training samples.

In most cases, the OPM method shows a more obvious enhancement than the OGM method. Based on the specific modulation design, the OPM method directly drops the feature of the better-performing modality in the feed-forward stage, allowing the suppressed modality to fully determine the multimodal prediction and temporarily control the optimization process. OGM weakens the gradient of the dominant modality in the back-propagation stage, slowing down the overall training. The modulation by OPM is stronger, which allows the suppressed modality to receive more targeted optimization efforts.

In addition, our methods show a more significant improvement on CREMA-D dataset compared to other datasets. The reason could be that data samples of CREMA-D dataset are recorded videos in controlled environments with less noise, while samples of other datasets are more noisy "in the wild" videos. Hence, for CREAM-D dataset, these clean samples are supposed to be easier to learn but suppressed by imbalanced multimodal learning. Then, after our modulation, multimodal model is more effectively enhanced.

4.3.2 Cross-modal interaction scenarios

Beyond the above vanilla and specifically designed fusion methods, various cross-modal interaction modules are proposed to improve the integration of different modalities in multimodal learning. In this section, we combine our methods with several multimodal models with cross-modal interaction modules, CentralNet [12], VATT [55] and MMTM [56], to evaluate their effectiveness in more complex cross-modal interaction scenarios. CentralNet [12] integrates the uni-modal feature of intermediate layers. VATT [55] uses the cross-modal attention mechanism. MMTM [56] activates the intermediate features of one modality with the guidance of others via the squeeze and excitation module, which

TABLE 6
Comparison with other dropout methods on CREMA-D, Kinetics-Sounds, and UCF-101 dataset. OPM and OGM methods are based on Concatenation fusion.

Method	CREMA-D		Kinetics-Sounds		UCF-101	
	Acc	mAP	Acc	mAP	Acc	mAP
Concatenation	66.9	72.1	63.0	67.3	80.5	86.4
ModDrop [31]	69.9	77.6	65.9	71.8	81.4	87.2
Dropout [27]	66.5	71.8	62.8	66.1	80.3	85.8
Annealed dropout [28]	65.3	70.9	62.1	65.2	80.1	84.7
Evolutional dropout [29]	65.9	72.0	62.4	65.7	80.4	86.1
Curriculum dropout [30]	64.9	70.2	61.9	64.9	79.9	84.3
OPM	**75.1**	81.2	**67.0**	**72.5**	**81.9**	**88.0**
OGM	74.6	**81.4**	65.4	71.4	81.5	87.6

TABLE 7
Accuracy of audio-visual event localization methods on AVE dataset, and audio-visual question answering methods on MUSIC-AVQA dataset.

Audio-visual Event Localization (Audio+Visual)				
Method	**w/o**	**w/ OPM**	**w/ OGM**	**w/ OPM+OGM**
AVGA [7]	73.6	75.6	**76.4**	76.2
AVSDN [19]	74.3	**75.1**	74.7	74.8
PSP [57]	74.4	**75.4**	75.3	75.3
Audio-visual Question Answering (Audio+Visual+Text)				
Method	**w/o**	**w/ OPM**	**w/ OGM**	**w/ OPM+OGM**
AVSD [58]	66.9	**67.3**	67.2	67.1
PanoAVQA [59]	70.3	70.8	71.0	**71.1**

can be regarded as self-attention on channels. Based on the results shown in Tab 3, our methods are not limited to the vanilla multimodal integration cases, but can also enhance the performance of multimodal models with more complex cross-modal interaction.

In addition, we also observe the change in discrepancy ratio during training of cross-modal interaction methods. According to Fig 5(a), CentralNet [12] method reduces the discriminative discrepancy between modalities, since it enhances the cross-modal interaction at the mid-level. Furthermore, our methods can be applied to these more complex scenarios, further reducing the discrepancy ratio and alleviating the imbalanced learning problem.

4.3.3 Comparison with uni-modal modulation methods

To address the issue that multimodal models often cannot jointly utilize all modalities effectively, uni-modal modulation methods have been proposed. To demonstrate the advantage of OPM and OGM, we make comparisons with other modulation methods: GBlending [8] and Greedy [11]. These methods either add uni-modal loss functions to the multimodal objective, controlling the training of individual modalities by loss weight [8], or enhance the optimization of other modalities by additional training step [11]. To ensure the fairness of experiments, the same backbone network and training settings are used. The comparison results are shown in Tab 5. We can have the following observations:

Firstly, GBlending [8] and Greedy [11] achieve better results than the baseline model with Concatenation fusion, which proves the effectiveness of these methods in solving the imbalanced under-optimized problem in multimodal models. Secondly, both our OPM and OGM methods achieve improvement compared with the Concatenation baseline. In particular, the OPM method obtains the best performance on the CREMA-D and UCF-101 datasets, outperforming other modulation methods. These results indicate the advantage of our on-the-fly modulation methods.

Moreover, it should be noted that GBlending [8] requires training additional uni-modal classifiers to provide a reference for modulation. Although it achieves the best accuracy on the Kinetics-Sounds dataset, it incurs higher training costs. As shown in Tab 4, the training time of GBlending [8] is considerably longer. Our methods are much more efficient while being effective.

4.3.4 Comparison with other dropout methods

In this section, we compare our methods with modality dropout method: ModDrop [31], and neuron dropout methods: Annealed dropout [28], Evolutional dropout [29] and Curriculum dropout [30]. ModDrop [31] drops modalities with a certain probability during training, easing the reliance of the multimodal model on any specific modality and improving model robustness. The results are shown in Tab 6. Based on the results, neuron dropout methods that drop neurons with a certain probability do not work well for the imbalanced multimodal learning problem because they focus mainly on neuron-level co-adaptation without considering modality-level discrepancies. In addition, although it is not designed for balanced multimodal learning, ModDrop [31] still has a gain in performance. As analyzed in Sec 3.1, the reason could be that the modality dropout provides each modality a chance to be optimized independently, getting rid of the suppression of others. Moreover, our methods are superior to these dropout methods, since they adaptively adjust the modulation of each modality during training by monitoring the discriminative discrepancy between modalities, considering the dynamical training process.

4.3.5 Complex multimodal task scenarios

The above experiments are either action recognition or emotion recognition, which are both video-level classifications. To further evaluate the proposed OPM and OGM methods in more general cases, we employ them on two complex multimodal tasks: audio-visual event localization and audio-visual question answering. The audio-visual event localization task aims to temporally demarcate both audible and visible events from a video. Therefore, it is in fact a segment-wise classification task. For each segment, there are predicted event labels and ground truth event labels. Global segment-wise classification accuracy is used as the evaluation metric in these experiments. The audio-visual question answering task involves selecting the correct answer for a given question about visual objects, sounds, and their associations. The widely-used AVE and MUSIC-AVQA datasets are used for these two tasks. Audio-visual event localization methods, AVGA [7], AVSDN [19] and PSP [57], are compared. Audio-visual question answering methods, AVSD [58] and PanoAVQA [59] are also compared.

The experiment results, as shown in Tab 7, indicate that both OPM and OGM maintain their effectiveness and en-

TABLE 8
Accuracy of emotion recognition task-oriented methods on the CREMA-D dataset.

Method	w/o	w/ OPM	w/ OGM
I-vector [60]	67.4	**75.2**	74.6
X-vector [61]	69.2	**74.7**	73.9
MWTSM [62]	68.4	**74.9**	74.1

TABLE 9
Combined with action recognition task-oriented methods on the Kinetics-Sounds and UCF-101 dataset.

Method	Kinetics-Sounds		UCF-101	
	Acc	mAP	Acc	mAP
TSN [63]	65.0	69.4	80.9	86.2
TSN-OPM	**70.0**	**75.9**	**82.5**	**87.7**
TSN-OGM	67.7	73.3	81.8	87.1
TSM [64]	66.0	70.2	81.2	86.5
TSM-OPM	**69.9**	**75.8**	82.3	87.5
TSM-OGM	67.5	73.0	**82.4**	**87.9**

hance performance in these more complex multimodal tasks. This demonstrates that the imbalanced learning phenomenon is not limited to video-level classification; it also affects more general multimodal cases. Moreover, to effectively capture the temporal correlations among modalities, both audio-visual event localization methods and audio-visual question answering methods often have more cross-modal interaction modules. Although this interaction design brings difficulty for the uni-modal performance estimation, our methods still obtain consistent performance gain. These experiments also indicate that our methods are effective on more complex multimodal tasks.

4.3.6 Combination with task-oriented methods

In this section, we combine OPM and OGM methods with several task-oriented methods to further evaluate their flexibility. Emotion recognition methods, I-vector [60], X-vector [61] and MWTSM [62] are compared on CREMA-D dataset. Action recognition methods, TSN [63] and TSM [64] are compared on Kinetics-Sounds and UCF-101 dataset. For all these methods, we utilize ResNet-18 as the backbone for both RGB and optical flow modalities and encoders are pre-trained on ImageNet on the UCF-101 dataset. In Tab 8 and Tab 9, we provide the experiment results. Both our OPM and OGM methods can be integrated with these specifically designed task-oriented models to further enhance performance. The experiments in combination with task-oriented methods further demonstrate the versatility of our OPM and OGM strategies on the more general jointly trained multimodal model.

4.3.7 Combined with multi-layer classifier case

In the analysis and methods section, we suppose that the final multimodal classifier is a single fully-connected layer without activation function, where the uni-modal discriminative performance could be estimated by splitting the weight of the final classifier. In fact, the core idea of our methods is to estimate the uni-modal prediction, and then use the estimated uni-modal prediction to evaluate the discriminative ability discrepancy between modalities.

TABLE 10
Combined with multi-layer classifier cases on CREMA-D, Kinetics-Sounds and UCF-101 dataset.

Method	CREMA-D		Kinetics-Sounds		UCF-101	
	Acc	mAP	Acc	mAP	Acc	mAP
Multi-layer classifier	67.7	73.2	63.7	68.4	80.3	85.2
Multi-layer classifier-OPM	**71.9**	**79.7**	**66.3**	**71.7**	**81.5**	**87.3**
Multi-layer classifier-OGM	69.3	77.9	64.5	70.3	81.2	87.0

TABLE 11
Experiments results on CMU-MOSI and UCF-101-Three datasets. Encoders of UCF-101-Three are pre-trained on ImageNet.

Method	CMU-MOSI (Audio+Visual+Text)		UCF-101-Three (RGB+OF+RGB Diff)	
	Acc	F1 score	Acc	mAP
Concatenation	75.9	76.0	82.5	88.6
OPM	**77.6**	**77.4**	**83.2**	89.0
OGM	76.8	76.7	82.8	88.6
OPM+OGM	77.1	76.7	83.2	**89.2**

Therefore, if the uni-modal prediction can be estimated, our methods are applicable beyond single-layer classifier scenarios. In this section, we consider the case that the final classifier is multi-layer. In such case, to estimate the uni-modal discriminative performance score s^m for modality m, we retain features of modality m while setting features of other modalities to 0, obtaining its uni-modal prediction. This zero-out strategy can be used for most multimodal cases. In similar studies [65], [66], this zero-out strategy has been widely used. For instance, Hu et al. [66] use zero-padding to replace modality x's feature to effectively simulate the multimodal prediction when modality x is absent. Hence, it is reasonable to utilize the zero-out strategy to simulate other modalities being absent, and estimate the prediction of remained modality. Based on the results in Tab 10, our methods maintain their efficacy. These results indicate that our idea of prediction modulation and gradient modulation could work well and extend to more complex classifier cases.

4.3.8 More-than-two modality cases

In this section, to further evaluate the effectiveness of our methods in scenarios involving more than two modalities, we conduct experiments on the CMU-MOSI and UCF-101-Three datasets. CMU-MOSI dataset covers three modalities, audio, visual and text. UCF-101-Three dataset has three modalities, RGB, optical flow, and RGB difference. These scenarios are more challenging than typical cases with only two modalities due to the more complex relationships among modalities. Based on the results in Tab 11, both OPM and OGM remain effective in these more complex multimodal scenarios. This demonstrates the scalability of our methods and their ability to adapt to diverse datasets.

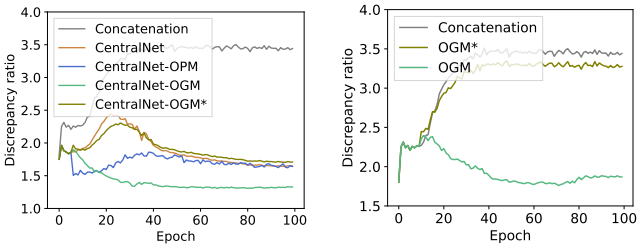
4.3.9 Combination of OPM and OGM methods

Our OPM and OGM methods are proposed to target the feed-forward stage and back-propagation stage, respectively. To explore whether these two methods can be combined, we conduct experiments on all used datasets. The results are shown in Tab 7, Tab 11 and Tab 12. It is observed that using

TABLE 12
Experiments about the combination of our OPM and OGM methods. All methods are based on Concatenation fusion.

Method	CREMA-D		Kinetics-Sounds	
	Acc	mAP	Acc	mAP
Concatenation	66.9	72.1	63.0	67.3
OPM	75.1	81.2	67.0	72.5
OGM	74.6	81.4	65.4	71.4
OPM+OGM	**76.7**	**83.6**	**68.0**	**74.7**

Method	UCF-101		VGGSound	
	Method	Acc	mAP	Acc
Concatenation	80.5	86.4	52.7	54.9
OPM	**81.9**	**88.0**	**54.1**	**56.5**
OGM	81.5	87.6	53.6	55.9
OPM+OGM	81.5	87.8	53.3	56.3



(a) (b)

Fig. 5. **Discrepancy ratio on Kinetics-Sounds dataset.** (a) Discrepancy ratio of Concatenation, CentralNet [12], and our methods during training. (b) Discrepancy ratio of Concatenation, OGM and OGM* methods based on Concatenation fusion during training. OGM* indicates the strategy that only increases the gradient of worse learnt modality. More discussion about OGM* is provided in Sec 4.4.2.

both OPM and OGM methods can yield better results. But this improvement is not consistent across all datasets. The reason could be that the estimation of uni-modal discriminative performance discrepancy is likely to be influenced when simultaneously conducting modulation at the feed-forward stage and back-propagation stage. For example, when dropping modalities first at the feed-forward stage, the modality discrepancy is changed. Then, the gradient modulation based on the original modality discrepancy is affected. Hence, how to well combine the modulation at both stages is expected to be further explored.

4.4 Fine-grained effectiveness analysis

4.4.1 Imbalance modulation analysis

In Fig 1, we present the uni-modal and multimodal performance after applying our methods. It is observed that the performance of both audio and visual encoders in the multimodal models is enhanced following our modulation. Additionally, an interesting observation is that our methods initially underperform compared to the baseline but ultimately surpass it. The reason could be that our modulation methods either drop the uni-modal feature or mitigate the gradient, decelerating the optimization of the modality with better performance at the beginning. And then the overall multimodal model exploits the information of another modality more and gains improvement in the end. To further analyze the modulation process, we monitor the change in discrepancy ratio between the two modalities during training on the Kinetics-Sounds dataset. The results are demonstrated in Fig 5. Based on the results, we can observe that the discrepancy ratio has an obvious drop after applying our modulation strategy, which provides concrete evidence for the effectiveness of our methods in alleviating the imbalanced learning problem. However, the two modalities are hard to have equal performance, *i.e.*, the discrepancy ratio is hard to be close to 1, since the uni-modal discriminative information is naturally not equal in the curated dataset.

4.4.2 Increase the gradient of worse learnt modality

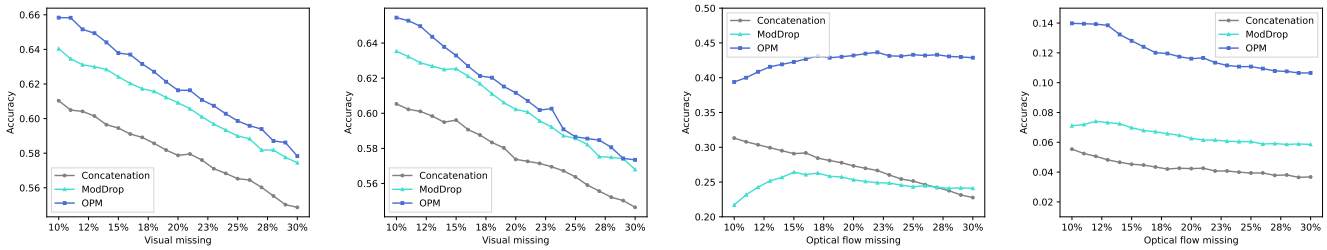
In OGM, we propose to mitigate the gradient of modality with better performance, and then the worse learnt modality can gain more training to ease the imbalanced learning problem. Corresponding to this idea of weakening the training of better modality, another idea is to accelerate the training of worse modality by increasing its gradient. To verify this opposite idea, we try the OGM* method, where the gradient of dominant modality remains unchanged, and the gradient of worse learnt modality is increased based on the uni-modal discrepancy. Experiments are conducted on CREMA-D and Kinetics-Sounds datasets. Based on the results in Tab 13, we can find that compared with the Concatenation fusion method, OGM* method can also improve the model performance, but it still has an obvious performance gap compared with the original OGM.

The reason could be that although OGM* accelerates the training of worse learnt modality by increasing its gradient, the training of more discriminative modality is not affected. More discriminative modality is still easy-to-learn for the multimodal model. Specifically, as shown in Fig 5(b), in OGM* method, the discriminative discrepancy ratio between modalities is still very large, which indicates the huge imbalance between modalities. The domination in multimodal optimization is not effectively weakened. Also, as shown in Fig 5(a), OGM* method is hard to alleviate the performance discrepancy with the more complex CentralNet multimodal model. Thus, OGM* only has a limited performance improvement. However, in OGM method, the training of more discriminative modality is directly suppressed by mitigating its gradient. OGM directly makes the more discriminative modality harder to learn than before, accordingly weakening its domination. As Fig 5(b), with OGM, the discriminative discrepancy ratio between modalities is greatly decreased, which indicates the imbalance between modalities is successfully alleviated. In Appendix F, we provide more experiments about the OGM* strategy.

These results demonstrate that increasing the gradient of worse learnt modality could accelerate the optimization of the corresponding modality. However, the increased gradient may be not suitable at both direction and scale in the global view, leading to limited enhancement.

4.4.3 Missing modality case of OPM

It is not always available data on all modalities during testing, which is a quite common problem in the practice. Since the OPM method uses the modality dropout strategy, it is hopeful to defend this problem to some extent since the model has been trained on the missing modality case during



(a) Audio missing=10% (b) Audio missing=20% (c) RGB missing=10% (d) RGB missing=20%

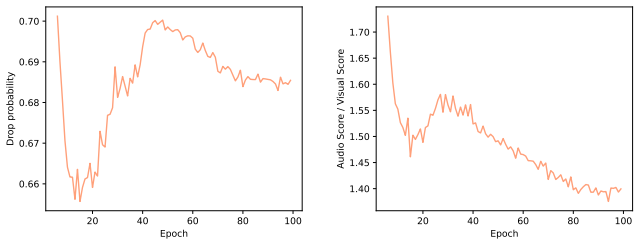
Fig. 6. **(a&b)**: Missing modality cases on Kinetics-Sounds dataset. **(c&d)**: Missing modality cases on UCF-101 dataset. OPM method is based on Concatenation fusion.

TABLE 13
Experiments about OGM and OGM* strategies on CREMA-D and Kinetics-Sounds dataset. OGM* aims to increase the gradient of suppressed modality. All methods are based on Concatenation fusion.

Method	CREMA-D		Kinetics-Sounds	
	Acc	mAP	Acc	mAP
Concatenation	66.9	72.1	63.0	67.3
OGM	**74.6**	**81.4**	**65.4**	**71.4**
OGM*	67.2	72.7	63.9	67.8

the training process. Here we conduct missing modality evaluation for a trained multimodal model. During evaluation, for a trained multimodal model, each modality of testing samples is randomly dropped by specific probability. For example, in Fig 6(a), for testing samples of Kinetics-Sounds dataset, the audio modality is dropped with a probability 10% and the visual modality is dropped with a probability ranging from 10% to 30%. In practice, data of the dropped modality is set to 0. The results are shown in Fig 6. Besides the Concatenation baseline, the ModDrop [31] method that drops each modality with a fixed probability during training is also compared. Based on the results, the idea of modality dropout indeed can overcome the missing modality scenario to some extent. Both OPM and ModDrop are superior to the Concatenation baseline in most cases. Besides, our OPM with adaptive modality dropout probability further brings the improvement in model robustness.

In addition, one observation is that the performance of ModDrop and OPM methods does not keep dropping when the missing probability of optical flow modality is increased, in Fig 6(c). This could be caused by the following reasons. For the UCF-101 dataset, RGB modality is clearly more discriminative than optical flow modality. As Fig 6(c) and Fig 6(d), when the missing probability of RGB modality increases from 10% to 20%, the accuracy of different methods drops dramatically. However, in contrast, as Fig 6(a) and Fig 6(b), when the missing probability of audio or visual modality increases, there is no such dramatic drop in accuracy. This demonstrates that the difference in discriminative ability between modalities on the UCF-101 dataset is more severe than others. For the UCF-101 dataset, multimodal models greatly depend on the more discriminative RGB modality even after modulation. But with the modulation of ModDrop and OPM methods, not only the ability of less discriminative optical flow modality, the ability of RGB modality in ModDrop and OPM is also further enhanced, compared with Concatenation baseline.



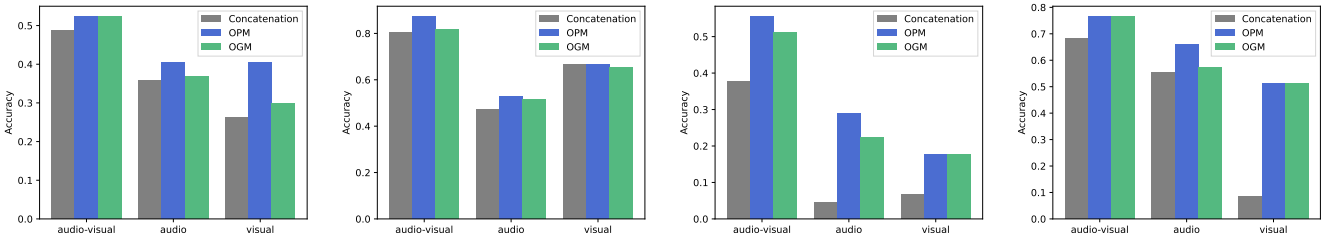
(a) (b)

Fig. 7. (a) Drop probability of audio modality of OPM on Kinetics-Sounds dataset. $q_{base} = 0.5$, $\lambda = 0.5$. The drop probability of visual modality remains 0. (b) Estimated discriminative score ratio of audio and visual modalities of OPM on Kinetics-Sounds dataset.

Therefore, when the missing probability of RGB modality is low (10% in Fig 6(c)), ModDrop and OPM methods with more powerful RGB encoders can still rely on the RGB modality to give correct predictions. Hence, their accuracy could even be marginally improved, when the noisy Optical-Flow modality is slightly absent. But as the missing percentage of Optical-Flow gradually increases, the accuracy of OPM and ModDrop begins to drop since too much Optical-Flow information is missing. When the missing probability of RGB modality is increased (20% in Fig 6(d)), even for ModDrop and OPM methods, models can not purely rely on RGB modality, and the significance of optical flow modalities begins to stand out. The performance experiences a continuous drop when the missing probability of optical flow modality increases. We provide more experiments about the influence of noisy modality on missing modality cases in Appendix E.

4.4.4 *Drop probability of each modality of OPM*

To further perceive the modulation of OPM during training, we record the specific changes in the modality dropout probability during training on the Kinetics-Sounds dataset. Results are shown in Fig 7(a). In our OPM method, the drop probability of the modality with more discriminative information is adaptively adjusted during training, based on the degree of discriminative discrepancy between modalities. Hence, the ratio of estimated discriminative scores (s in Eqn 6) between audio and visual modalities are also recorded and shown in Fig 7(b). Based on the results, the discriminative discrepancy between modalities changes dynamically during the training process, rather than remaining static. In addition, it can be seen that audio, as the dominant modality, always owns a higher accuracy than visual modality in the training process. Hence, the dropout probability for the audio modality is dynamic, while for the



(a) playing saxophone (b) dribbling basketball (c) bathroom fan running (d) telephone bell ringing

Fig. 8. **(a&b)**: Category-wise analysis on Kinetics-Sounds dataset. **(c&d)**: Category-wise analysis on VGGSound dataset. OPM and OGM methods are based on Concatenation fusion.

visual modality it remains at 0 throughout training. The change in the estimated discriminative score ratio of audio and visual modalities reflects the need to adaptively adjust the dropout probabilities of each modality during training.

4.4.5 Category-wise analysis

To provide a fine-grained analysis of our methods, we conduct a category-wise analysis on the Kinetics-Sounds and VGGSound datasets. The results are shown in Fig 8. Both our OPM and OGM methods improve the performance category-wise to a certain degree. Furthermore, we notice that the modality with less discriminative information generally achieves greater performance improvements when utilizing our methods. For example, the *playing saxophone* category is more discriminative in audio and its audio modality earns less improvement. This phenomenon is consistent with our modulation which mitigates the influence of the dominant modality and then improves the learning of worse learnt modality. In addition, it is also validated that regardless of which modality dominates the training for a category, the OPM and OGM methods can effectively alleviate the imbalance. For instance, the *dribbling basketball* and *bathroom fan running* categories are more discriminative in visual (gray bar for audio is lower than visual in Fig 8(b) and Fig 8(c)), while the *playing saxophone* and *telephone bell ringing* categories have a preference in audio (gray bar for audio is higher than visual in Fig 8(a) and Fig 8(d)), and they both obtain enhancement.

Furthermore, although VGGSond is primarily a sound-oriented dataset with categories mainly determined by sound type, this does not mean that the visual information is non-discriminative. While the visual information is not directly related to the "auditory label," it does reflect the corresponding sound events through the objects producing those sounds. The visual information of sounding objects can aid in distinguishing one sound type from another. Therefore, when our methods enhance the learning of the visual modality, the model's classification ability improves due to the strengthened visual capability.

4.5 Ablation study

4.5.1 Other variants of our modulation strategies

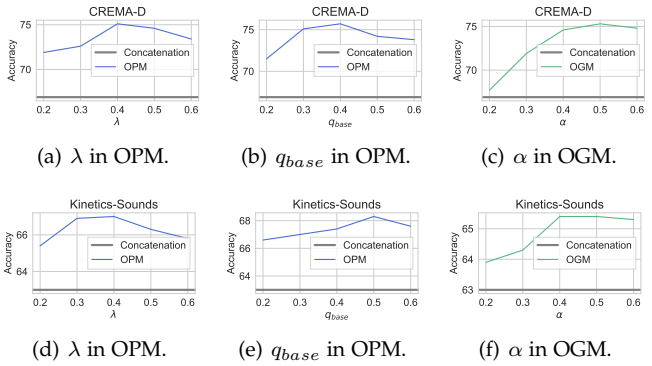
Besides the strategy proposed in the method section, we also attempt several alternative strategies in this section. Notion follows Sec 3.2. Firstly, as stated in Sec 3.2, s^m is the predicted probability for the correct category of modality m. Then, the total discriminative score of modality m in a mini-batch is $\sum_{i\in B_t} s_i^m$. In Sec 3.2, the discriminative discrepancy degree among modalities is measured as the average ratio of this score, i.e., $\rho_t^m = \frac{1}{M-1}\sum_{j\in[M],j\neq m}\frac{\sum_{i\in B_t} s_i^m}{\sum_{i\in B_t} s_i^j}$. Here, we attempt to measure the discriminative discrepancy degree via the difference of this score, i.e., $\rho_t^m = \frac{1}{M-1}\sum_{j\in[M],j\neq m}(\sum_{i\in B_t} s_i^m - \sum_{i\in B_t} s_i^j)$. Secondly, in the above experiments, $tanh(x-1)$ is used as the monotonically increasing function $z(x)$, which is used to map the discriminative discrepancy degree between modalities (*i.e.*, ρ_t) into $(0,1)$ as the strength of modulation. We further attempt to use $sigmoid(x)$ function as $z(x)$ in this section. The results are shown in Tab 15. We can have several observations. Firstly, the measurement of discriminative discrepancy degree between modalities does not greatly affect the effectiveness. Both the ratio of scores and the

TABLE 14
Different split of bias term in the final classifier. All methods are based on Concatenation fusion.

Dataset	CREMA-D		Kinetics-Sounds	
Concatenation	66.9		63.0	
Method	OPM	OGM	OPM	OGM
$\gamma^1=0,\gamma^2=0$	74.5	73.9	66.9	65.4
$\gamma^1=1,\gamma^2=1$	74.2	74.7	66.7	64.9
$\gamma^1=\frac{1}{2},\gamma^2=\frac{1}{2}$	75.1	74.6	67.0	65.4
$\gamma^1=\frac{1}{3},\gamma^2=\frac{2}{3}$	74.6	74.3	66.7	65.1
$\gamma^1=\frac{2}{3},\gamma^2=\frac{1}{3}$	73.7	74.5	66.5	64.7

TABLE 15
Accuracy of the variants of our modulation strategies on the Kinetics-Sounds and UCF-101 dataset. OPM and OGM methods are based on Concatenation fusion. KS denotes for Kinetics-Sounds.

Dataset	KS		UCF-101	
Concatenation	63.0		80.5	
Method	OPM	OGM	OPM	OGM
$\rho_t^m = \frac{1}{M-1}\sum_{j\in[M],j\neq m}\frac{\sum_{i\in B_t} s_i^m}{\sum_{i\in B_t} s_i^j}$, $z(x)=tanh(x-1)$	67.0	65.4	81.9	81.5
$\rho_t^m = \frac{1}{M-1}\sum_{j\in[M],j\neq m}\frac{\sum_{i\in B_t} s_i^m}{\sum_{i\in B_t} s_i^j}$, $z(x)=sigmoid(x)$	67.2	64.2	81.7	80.9
$\rho_t^m = \frac{1}{M-1}\sum_{j\in[M],j\neq m}(\sum_{i\in B_t} s_i^m - \sum_{i\in B_t} s_i^j)$, $z(x)=tanh(x-1)$	66.9	65.4	81.8	80.8
$\rho_t^m = \frac{1}{M-1}\sum_{j\in[M],j\neq m}(\sum_{i\in B_t} s_i^m - \sum_{i\in B_t} s_i^j)$, $z(x)=sigmoid(x)$	65.8	65.4	81.8	80.9



(a) λ in OPM. (b) q_{base} in OPM. (c) α in OGM.

(d) λ in OPM. (e) q_{base} in OPM. (f) α in OGM.

Fig. 9. Hyper-parameter sensitivity analysis of OPM and OGM on CREMA-D, Kinetics-Sounds dataset. OPM and OGM methods are based on Concatenation fusion.

difference of scores can be used as the measure. Secondly, it does not need much effort on the specific selection of $z(x)$. Different combinations of ρ_t and $z(x)$ can obtain consistency improvement compared with the Concatenation baseline across different datasets. Overall, these experiments indicate the effectiveness of the proposed modulation idea, demonstrating that our methods are not dependent on a specific design.

4.5.2 Split of bias term in the final classifier

In Eqn 6, when estimating the uni-modal discriminative performance for modality m, we use $W_i^m \cdot \varphi_i(\theta^m, x_i^m) + \frac{b}{M}$. The bias term in the final classifier is equally split. In this section, we try different splits of bias term: $W_i^m \cdot \varphi_i(\theta^m, x_i^m) + \gamma^m \cdot b$. Based on the results in Tab 14, different splits of this bias term in the final classifier do not have a great influence on the effectiveness of our methods. The reason could be that the value of bias term is smaller than $W_i^m \cdot \varphi_i(\theta^m, x_i^m)$. Therefore, its split does not affect the uni-modal discriminative performance estimation a lot. Then, the effectiveness of our method has no reliance on the split of this bias term.

4.5.3 Hyper-parameter sensitivity analysis

In this section, we provide the hyper-parameter sensitivity analysis on the CREMA-D, Kinetics-Sounds and UCF-101 datasets. We select different values of λ and q_{base} in OPM as well as α in OGM. q_{base} controls the lower bound and upper bound of modality dropout probability in OPM, while λ and α control the degree of the modulation. The results are shown in Fig 9. According to the results, although the value of hyper-parameters with the best performance varies on different datasets, all selections of these three hyper-parameters can obtain consistent improvement compared with the Concatenation baseline. Therefore, it does not need much effort on the selection of hyper-parameters.

5 Conclusion

In this paper, we first observe and analyze the imbalanced learning phenomenon from both the feed-forward and the back-propagation stage in the multimodal learning, then propose on-the-fly modulation methods, OPM and OGM, to alleviate this problem. OPM mitigates the influence of the dominant modality by dropping its feature with dynamical probability in the feed-forward stage, while OGM weakens its gradient in the back-propagation stage. Both OPM and OGM aim to help the suppressed modality obtain more training. Moreover, our modulation methods are expected to have more sufficient learning of multimodal representation, since it enhances the learning of each modality. In the experiment, we combine the proposed methods with different multimodal models on various tasks and our methods achieve considerable improvement. A wide range of fine-grained analyses, several alternative variants and hyper-parameter sensitivity analyses are also provided to perceive their effectiveness from different views. Extensive experiment results demonstrate the promising effectiveness and versatility of the proposed methods.

Acknowledgments

Zequn Yang participated in the theoretical analysis discussion. We deeply thank him for his valuable assistance. This work was supported by National Natural Science Foundation of China (NO.62106272), the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), and Public Computing Cloud, Renmin University of China.

References

[1] E. B. Herreras, "Cognitive neuroscience; the biology of the mind," *Cuadernos de Neuropsicología/Panamerican Journal of Neuropsychology*, vol. 4, no. 1, pp. 87–90, 2010.
[2] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.
[3] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019.
[4] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *CVPR*, 2020.
[5] A. R. Choudhury, R. Vanguri, S. R. Jambawalikar, and P. Kumar, "Segmentation of brain tumors using deeplabv3+," in *International MICCAI Brainlesion Workshop*, 2018.
[6] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation," *arXiv preprint arXiv:2108.10528*, 2021.
[7] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.
[8] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*, 2020.
[9] Y. Sun, S. Mai, and H. Hu, "Learning to balance the learning rates between various modalities via adaptive tracking factor," *IEEE Signal Processing Letters*, vol. 28, pp. 1650–1654, 2021.
[10] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably)," *arXiv preprint arXiv:2203.12221*, 2022.
[11] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *ICML*, 2022.
[12] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *ECCV Workshops*, 2018.
[13] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.
[14] K. Parida, N. Matiyali, T. Guha, and G. Sharma, "Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos," in *WACV*, 2020.
[15] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022.
[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *NeurIPS*, 2014.

[17] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

[18] D. Hu, X. Li *et al.*, "Temporal multimodal learning in audiovisual speech recognition," in *CVPR*, 2016.

[19] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *ICASSP*, 2019.

[20] I. Ilievski and J. Feng, "Multimodal learning and reasoning for visual question answering," in *NeurIPS*, 2017.

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

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

[23] Y. Wei and D. Hu, "Mmpareto: boosting multimodal learning with innocent unimodal assistance," in *ICML*, 2024.

[24] Y. Wei, R. Feng, Z. Wang, and D. Hu, "Enhancing multimodal cooperation via sample-level modality valuation," in *CVPR*, 2024.

[25] Y. Wei, S. Li, R. Feng, and D. Hu, "Diagnosing and re-learning for balanced multimodal learning," in *ECCV*, 2024.

[26] Z. Yang, Y. Wei, C. Liang, and D. Hu, "Quantifying and enhancing multi-modal robustness with modality preference," *arXiv preprint arXiv:2402.06244*, 2024.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] S. J. Rennie, V. Goel, and S. Thomas, "Annealed dropout training of deep networks," in *SLT Workshop*, 2014.

[29] Z. Li, B. Gong, and T. Yang, "Improved dropout for shallow and deep learning," *NeurIPS*, 2016.

[30] P. Morerio, J. Cavazza, R. Volpi, R. Vidal, and V. Murino, "Curriculum dropout," in *ICCV*, 2017.

[31] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[32] X. Li, Q. Dou, H. Chen, C.-W. Fu, and P.-A. Heng, "Multi-scale and modality dropout learning for intervertebral disc localization and segmentation," in *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, 2016, pp. 85–91.

[33] A. Hussen Abdelaziz, B.-J. Theobald, P. Dixon, R. Knothe, N. Apostoloff, and S. Kajareker, "Modality dropout for improved performance-driven talking faces," in *ICMI*, 2020.

[34] S. de Blois, M. Garon, C. Gagné, and J.-F. Lalonde, "Input dropout for spatially aligned modalities," in *ICIP*, 2020.

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

[36] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma, "The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects," in *ICML*, 2019.

[37] P. Chaudhari and S. Soatto, "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks," in *ITA Workshop*, 2018.

[38] Z. Xie, F. He, S. Fu, I. Sato, D. Tao, and M. Sugiyama, "Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting," *Neural computation*, vol. 33, no. 8, pp. 2163–2192, 2021.

[39] J. Wu, W. Hu, H. Xiong, J. Huan, V. Braverman, and Z. Zhu, "On the noisy gradient descent that generalizes as sgd," in *ICML*, 2020.

[40] F. He, T. Liu, and D. Tao, "Control batch size and learning rate to generalize well: Theoretical and empirical evidence," *NeurIPS*, 2019.

[41] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, "Three factors influencing minima in sgd," *arXiv preprint arXiv:1711.04623*, 2017.

[42] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *ICML*, 2017.

[43] A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens, "Adding gradient noise improves learning for very deep networks," *stat*, vol. 1050, p. 21, 2015.

[44] M. Zhou, T. Liu, Y. Li, D. Lin, E. Zhou, and T. Zhao, "Toward understanding the importance of noise in training neural networks," in *ICML*, 2019.

[45] C. Wei, S. Kakade, and T. Ma, "The implicit and explicit regularization effects of dropout," in *ICML*, 2020.

[46] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

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

[48] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *CVPR*, 2017.

[49] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[50] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[51] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[52] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *CVPR*, 2022.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[54] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *ECCV*, 2018.

[55] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *NeurIPS*, 2021.

[56] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *CVPR*, 2020.

[57] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in *CVPR*, 2021.

[58] I. Schwartz, A. G. Schwing, and T. Hazan, "A simple baseline for audio-visual scene-aware dialog," in *CVPR*, 2019.

[59] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in *ICCV*, 2021.

[60] P. Heracleous and A. Yoneyama, "A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme," *PloS one*, vol. 14, no. 8, p. e0220386, 2019.

[61] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP*, 2020.

[62] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *ACII*, 2019.

[63] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[64] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019.

[65] A. Ghorbani and J. Y. Zou, "Neuron shapley: Discovering the responsible neurons," *NeurIPS*, 2020.

[66] P. Hu, X. Li, and Y. Zhou, "Shape: An unified approach to evaluate the contribution and cooperation of individual modalities," *IJCAI*, 2022.