Interactive Multimodal Learning via Flat Gradient Modification

Qing-Yuan Jiang, **Zhouyang Chi**, **Yang Yang***
Nanjing University of Science and Technology
{jiangqy, yyang}@njust.edu.cn, zhouyangchi0@gmail.com

Abstract

Due to the notorious modality imbalance phenomenon, multimodal learning (MML) struggles to achieve satisfactory performance. Recently, multimodal learning with alternating unimodal adaptation (MLA) has been proven effective in mitigating the interference between modalities by capturing interaction through orthogonal projection, thus relieving modality imbalance phenomenon to some extent. However, the projection strategy orthogonal to the original space can lead to poor plasticity as the alternating learning proceeds, thus affecting model performance. To address this issue, in this paper, we propose a novel multimodal learning method called interactive MML via flat gradient modification (IGM) by employing a flat gradient modification strategy to enhance interactive MML. Specifically, we first employ a flat projection-based gradient modification strategy that is independent to the original space, aiming to avoid the poor plasticity issue. Then we introduce the sharpness-aware minimization (SAM)-based optimization strategy to fully exploit the flatness of the learning objective and further enhance interaction during learning. To this end, the plasticity problem can be avoided and the overall performance is improved. Extensive experiments on widely used datasets demonstrate that IGM outperforms various state-of-the-art (SOTA) baselines, achieving superior performance. The source code is available at https://github.com/njustkmg/IJCAI25-IGM.

1 Introduction

Multimodal learning (MML) [Zhao *et al.*, 2016; Perez *et al.*, 2018; Yang *et al.*, 2019; Li *et al.*, 2020; Du *et al.*, 2022; Liang *et al.*, 2022] has attracted much attention and made promising progress across a wide range of real applications such as speech recognition [Ngiam *et al.*, 2011], sentiment analysis [Zhu *et al.*, 2024], image caption [Chang *et al.*, 2015], multimedia retrieval [Wang *et al.*, 2016; Zhu *et al.*, 2023; Yang *et al.*, 2024a], recommendation system [Xiao *et al.*, 2022], and so on.

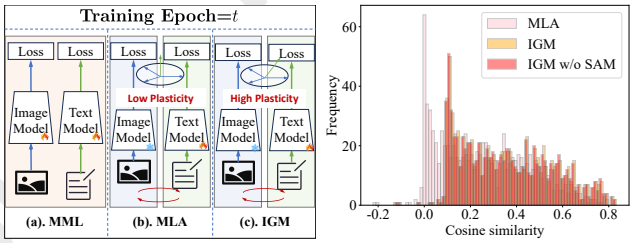


Figure 1: Illustration of motivation. **Left:** The architecture comparison for MML, MLA and IGM. **Right:** The histgram of gradient change cosine similarity for MLA, IGM w/o SAM, and IGM.

Compared with the single modality method, multimodal learning methods are expected to achieve better performance through fusing rich information from multiple modalities. However, as the information among the different modalities is imbalanced, multimodal learning, which usually tries to optimize a uniform objective, falls into the trap of focusing on learning the dominant modality while ignoring the non-dominant modality [Wang *et al.*, 2020; Peng *et al.*, 2022]. Hence, the overall performance of multimodal learning in practical applications is greatly restricted because of the modality imbalance problem.

In recent years, many works [Wang *et al.*, 2020; Zong *et al.*, 2024; Yang *et al.*, 2025; Yang *et al.*, 2024b] have explored the modality imbalance problem and various algorithms have been proposed to balance the learning procedure for all modalities. The paradigm of these methods is illustrated in the left Sub-Figure 1 (a), where late fusion is used as an example for general MML. Among these methods, most of them [Wang *et al.*, 2020; Peng *et al.*, 2022; Fan *et al.*, 2023; Li *et al.*, 2023] focus on designing a learning adjustment strategy to rebalance the learning speeds for different modalities. Other representative methods [Wu *et al.*, 2022a; Du *et al.*, 2023] introduce extra networks as the auxiliary module to overcome the modality imbalance problem. Unfortunately, these methods usually optimize the multiple modality-specific models simultaneously, thus failing to fully explore the interaction between all modalities and affecting the model performance. Recent work MLA [Zhang *et al.*, 2024] designs an alternating unimodal adaption algorithm to

*Corresponding author

capture the cross-modality information. As shown in the left Sub-Figure 1 (b), MLA employs the orthogonal projection to capture the cross-modal interaction, thus mitigating the interference between different modalities and further relieving the modality imbalance phenomenon. However, [Zhao *et al.*, 2023] finds that the orthogonal projection strategy used to transfer gradient information to promote learning leads to poor plasticity problem. For MLA, as the alternating learning proceeds, the influence of orthogonal projection is continuously imposed on the model, leading to feasible gradient direction becomes narrow, i.e., poor plasticity. This issue results in a suboptimal solution.

To address this issue, we propose a novel flat projection-based gradient modification (GM) strategy to facilitate the capturing of cross-modal interactions. Essentially, flatness and sharpness [Chaudhari *et al.*, 2017; Keskar *et al.*, 2017] characterize the nature of the loss landscape. The flatter the direction of gradient transfer, the more the original modal information can be preserved due to the stability of the flat direction. More importantly, because the selection of the flat direction is based on the loss of the current modality and independent of the gradient direction of the affected modality, this strategy avoids poor plasticity issue. In summary, the flat projection-based strategy can mitigating the interference between different modalities and address the poor plasticity problem simultaneously. As plasticity [Sun *et al.*, 2022] refers to a model's ability to adapt to new modality after learning previous modality, we compare the gradient change histgram by calculating the cosine similarity between the gradients of the old and new modalities in the right Sub-Figure 1. The results demonstrate that the flat projection-based gradient modification method (IGM w/o SAM) achieves higher similarity compared to MLA, suggesting that this strategy effectively mitigates the poor plasticity issue. In addition, the issue of poor plasticity was further confirmed through accuracy comparisons in Table 1. Furthermore, because well-known highly non-convex [Foret *et al.*, 2021; Deng *et al.*, 2021] of the loss of deep neural networks, the loss landscape is usually sharp. To further explore and employ the flatness of the loss landscape in multimodal learning, we introduce the SAM-based [Foret *et al.*, 2021] optimization strategy to smooth the learning objective. By introducing this strategy, we strengthen the flatness of the loss landscape and thus enhance the effectiveness of interactive learning. These two novel strategies are illustrated in the left Sub-Figure 1 (c).

Our proposed novel approach is named as interactive MML via flat gradient modification (IGM). Our contributions can be summarized as follows:

- We propose a novel flat projection-based gradient modification strategy to capture the cross-modal interaction. This strategy can avoid the poor plasticity caused by orthogonal projection.
- To further employ the flatness of the loss landscape, we introduce a SAM based optimization strategy to smooth the learning objective.
- Extensive experiments on widely used datasets show that our IGM can outperform state-of-the-art baselines to achieve the best performance.

2 Related Works

2.1 Imbalance Multimodal Learning

Because of modality imbalance, MML methods sometimes exhibit the counterintuitive phenomenon of performing worse than unimodal models [Peng *et al.*, 2022]. Due to the heterogeneity, different models converge at different rates during training, leading to suboptimal performance in MML.

Some researchers have proposed a serious of approaches [Wang *et al.*, 2020; Peng *et al.*, 2022; Fan *et al.*, 2023; Li *et al.*, 2023; Wei and Hu, 2024] to address this problem by rebalancing the modal learning. To be more specific, these approaches aim to slow down the learning of dominant modality by adjusting the gradients to ensure that the learning of both modalities is as balanced as possible. Other attempts, including uni-modal teacher (UMT) [Du *et al.*, 2023] and greedy MML [Wu *et al.*, 2022b], employ an extra network module to assist MML. Both methods adopt a learning paradigm that updates the parameters of all modalities simultaneously. To enhance the interaction among all modalities, MLA [Zhang *et al.*, 2024] employs an alternating learning paradigm for interactive MML, which leads to performance improvement.

2.2 Sharpness Aware Minimization

Many efforts have been made to overcome the highly non-convex problem of DNN models by using the properties of the loss landscape. Sharpness aware minimization (SAM) [Foret *et al.*, 2021] proposes an effective algorithm to improve the generalization ability by using the relationship between loss sharpness and generalization. In particular, instead of learning the original objective, SAM aims to minimize the loss value and loss sharpness simultaneously. The learned parameters by SAM usually lie in the neighborhoods that have uniformly loss value of the original objective, leading to converging to flat minima. Therefore, the loss landscape will be more flat and the objective will converge to a flat minimum. SAM has been applied in many application scenarios successfully. For example, FlatMatch [Huang *et al.*, 2023] extends SAM to semi-supervised learning by penalizing the cross-sharpness between the worst-case model and the original model.

3 Methodology

In this paper, we focus on late fusion MML approach which usually adopt a two stream architecture following the setting of [Ye *et al.*, 2018; Liu *et al.*, 2024]. We present our proposed multimodal representation learning method IGM in detail. The whole IGM approach is shown in Figure 2. IGM contains two important components, i.e., flat projection-based gradient modification and SAM-based optimization.

3.1 Preliminary

Assume that we have n data entities for training, each of which contains m modalities. Without loss of generality, we use $\mathcal{D} = \{\mathcal{X}^{(j)}\}_{j=1}^{m}$ to denote the training set, where $\mathcal{X}^{(j)} = \{\boldsymbol{x}_i^{(j)}\}_{i=1}^{n}$ denotes the data points of j-th modality and $\boldsymbol{x}_i^{(j)}$ denotes the i-th data point. In addition, we are also given a category label $\boldsymbol{y}_i \in \{0,1\}^c$ for each data point, where

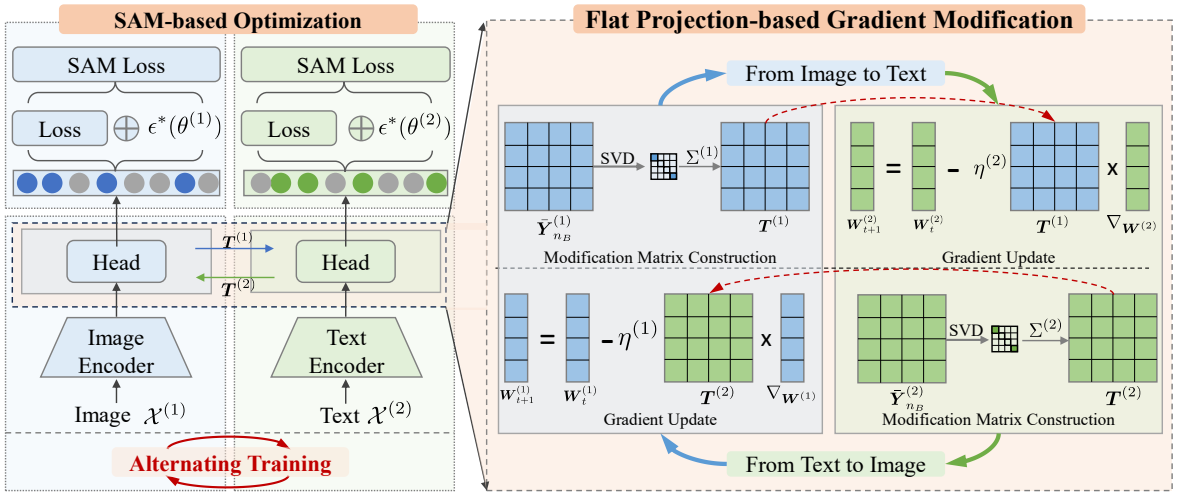


Figure 2: The architecture of our proposed IGM. Our method contains two key components, i.e., flat projection-based gradient modification (shown in the right part of the panel) and SAM-based optimization (shown in the upper left corner of the panel).

c denotes the number of category labels. In general, the goal of multimodal learning is to use the training set $\mathcal{D}$ to learn a model to predict category labels for unseen data.

For deep multimodal learning methods [Wang *et al.*, 2020; Peng *et al.*, 2022; Li *et al.*, 2023], different deep neural networks are used as the model to predict categories for each modality. For the sake of simplicity, we use $\varphi^{(j)}(\cdot)$ to denote the encoder which is used to extract the feature of j-th modality. And the feature $\boldsymbol{z}_i^{(j)}$ of i-th data point can be calculated by $\boldsymbol{z}_i^{(j)} = \varphi^{(j)}(\boldsymbol{x}_i^{(j)}; \Phi^{(j)})$, where $\Phi^{(j)}$ denotes the parameters. Then the prediction $\boldsymbol{p}_i^{(j)}$ can be presented as:

$$\boldsymbol{p}_i^{(j)} = \phi^{(j)}(\boldsymbol{z}_i^{(j)}; \Theta^{(j)}) = \texttt{softmax}([\boldsymbol{W}^{(j)}]^\top \boldsymbol{z}_i^{(j)}),$$

where $\phi^{(j)}(\cdot)$ indicates the j-th classifier, $\Theta^{(j)}$ denotes the parameter of j-th classifier, and $\boldsymbol{W}^{(j)}$ denotes the weight of fully-connected layer. According to $\boldsymbol{p}_i^{(j)}$, the training procedure is performed by minimizing the following loss function:

$$L(\theta^{(j)}; \mathcal{X}^{(j)}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i^\top \log \boldsymbol{p}_i^{(j)}, \tag{1}$$

where $\theta^{(j)} \triangleq \{\Phi^{(j)}, \Theta^{(j)}\}$. After training, the final prediction of i-th data point can be generated by the following equation:

$$\boldsymbol{p}_i = f(\boldsymbol{p}_i^{(1)}, \cdots, \boldsymbol{p}_i^{(m)}).$$

Here, $f(\cdot)$ denotes the late fusion strategy. In practice, there exist various late fusion strategies like averaging or weighting. However, how to design fusion strategies is not the focus of our paper and it will be left for future study.

Unlike to most of late fusion approaches, MLA tries to establish the connections between the learning processes in different modalities. Specifically, MLA designs an alternating learning paradigm to capture the interaction through orthogonal projection. However, MLA adopts an orthogonal projection strategy to perform interactive learning and usually suffers from poor plasticity problem [Wang *et al.*, 2021] due to the usage of orthogonal projection strategy.

3.2 Flat Projection-based Gradient Modification

Flatness and sharpness [Chaudhari *et al.*, 2017; Keskar *et al.*, 2017] are two pivotal properties of loss. The change of loss value is relatively smooth in the flat directions. Hence, when we transfer the gradient information along with the flat directions, the information we want to transfer will be less affected by the change of loss.

Then, inspired by Adam-NSCL [Wang *et al.*, 2021], we design a singular value decomposition (SVD) based approach to find the flat directions. We use the training procedure of k-th and l-th modality to illustrate the flat projection-based gradient modification strategy. We utilize the full-connected layer before the classification layer to illustrate the flat direction modification strategy. Given t-th batch of n_B samples $\mathcal{X}_t^{(k)} = \{\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_{n_B}^{(k)}\}$, the features of the input batch can be calculated by:

$$\boldsymbol{Z}_t^{(k)} = \varphi^{(k)}(\mathcal{X}_t^{(k)}),$$

where $\boldsymbol{Z}_t^{(k)} \in \mathcal{R}^{n_B \times d}$, d is the dimensionality of feature. Then we compute the mean of features and the covariance of the batch by:

$$\bar{\boldsymbol{z}}_t^{(k)} = \texttt{mean}(\boldsymbol{Z}_t^{(k)}) \in \mathcal{R}^d,$$
$$\boldsymbol{Y}_t^{(k)} = \bar{\boldsymbol{z}}_t^{(k)}[\bar{\boldsymbol{z}}_t^{(k)}]^\top \in \mathcal{R}^{d \times d}.$$

Then, the cumulative variance can be calculated by:

$$\begin{cases} \bar{\boldsymbol{Y}}_t^{(k)} = \boldsymbol{Y}_t^{(k)}, & \text{if } t = 1, \\ \bar{\boldsymbol{Y}}_t^{(k)} = \bar{\boldsymbol{Y}}_{t-1}^{(k)} + \boldsymbol{Y}_t^{(k)}, & \text{otherwise.} \end{cases} \tag{2}$$

By applying SVD to $\bar{\boldsymbol{Y}}_{n_B}^{(k)}$, we have:

$$\boldsymbol{U}^{(k)}\Lambda^{(k)}[\boldsymbol{V}^{(k)}]^\top \triangleq \texttt{svd}(\bar{\boldsymbol{Y}}_{n_B}^{(k)}),$$

where $\Lambda^{(k)} = \texttt{diag}(\lambda_1^{(k)}, \cdots, \lambda_d^{(k)})$ denotes the singular values matrix, $\boldsymbol{U}^{(k)}$ and $\boldsymbol{V}^{(k)} = [\boldsymbol{v}_1^{(k)}, \cdots, \boldsymbol{v}_d^{(k)}]$ denote the left and right singular vectors, respectively.

For now, let us consider the geometry properties of the direction indicated by singular vector $\boldsymbol{v}_i^{(k)}$. If we perturb $\boldsymbol{Z}_t^{(k)}$ along with the singular direction $\boldsymbol{v}_i^{(k)}$ with the perturbation magnitude $\gamma \boldsymbol{v}_i^{(k)}$, the change of the output for the last full-connected layer can be computed by:

$$\|\boldsymbol{Z}_t^{(k)} \gamma \boldsymbol{v}_i^{(k)}\| = \|\boldsymbol{U}^{(k)} \Lambda^{(k)} [\boldsymbol{V}^{(k)}]^\top \gamma \boldsymbol{v}_i^{(k)}\| = \gamma \lambda_i^{(k)}. \quad (3)$$

From Equation (3), the flatness of the direction indicated by singular vector $\boldsymbol{v}_i^{(k)}$ is determined by the singular value $\lambda_i^{(k)}$. In other words, the larger the singular value, the smaller the update modification should be in the direction of the singular vector. Thus, we design the following gradient modification matrix $\boldsymbol{T}^{(k)}$:

$$\boldsymbol{T}^{(k)} = \boldsymbol{V}^{(k)} \Sigma^{(k)} [\boldsymbol{V}^{(k)}]^\top. \quad (4)$$

Here, $\Sigma^{(k)} = \exp\big(-\frac{\tau}{\lambda_{\max}^{(k)} - \lambda_{\min}^{(k)}} (\Lambda^{(k)} - \lambda_{\min}^{(k)} \boldsymbol{I})\big)$, $\tau > 0$ is a scaling factor and $\boldsymbol{I}$ is the identity matrix. Hence, when we update the parameter of l-th modality for the last full-connected layer, the SGD-based update rule is modified as:

$$\boldsymbol{W}_{t+1}^{(l)} = \boldsymbol{W}_t^{(l)} - \eta^{(l)} \boldsymbol{T}^{(k)} \nabla_{\boldsymbol{W}^{(l)}} L(\theta^{(l)}), \quad (5)$$

where $\eta^{(l)}$ denotes the corresponding learning rate. From Equation (5), we can find that the information of k-th modality is injected into l-th modality. Equipped with the flat projection-based gradient modification strategy, we can transfer the original modality information more effectively with less impact. From Equation (4), the calculation of gradient modification matrix is independent of the l-th modality. Hence, our strategy can avoid poor plasticity problem.

3.3 SAM-based Optimization

Up to now, we design a novel gradient projection strategy to address the poor plasticity issue. Unfortunately, the loss of DNN is usually highly non-convex, i.e., the loss landscape is usually sharp. To better find a flat direction, we introduce the SAM [Foret *et al.*, 2021]-based optimization strategy to smooth the learning objective.

Because the SAM-based optimization strategy will be applied to all modalities, we omit the superscript "(k)" and use θ directly to illustrate. For multimodal learning with loss $L(\theta)$, we define the perturbation of parameter θ as ϵ. Based on ϵ, the SAM objective [Foret *et al.*, 2021; Huang *et al.*, 2023] can be defined as:

$$\begin{aligned} L^{\text{SAM}}(\theta) &\triangleq \max_{\epsilon:\|\epsilon\|_p \le \rho} L(\theta + \epsilon), \\ &\triangleq \max_{\epsilon:\|\epsilon\|_p \le \rho} \frac{1}{n} \sum\nolimits_{i=1}^{n} \ell(\theta + \epsilon; \boldsymbol{p}_i, \boldsymbol{y}_i), \end{aligned}$$

where ρ restricts the perturbation magnitude of θ within ℓ_p-ball. Instead of minimizing the objective function $L(\theta)$ in Equation (1), we perturb the parameter θ with $\epsilon \in \Psi$ and optimize the following SAM objective:

$$\min_\theta L^{\text{SAM}}(\theta). \quad (6)$$

Here, Ψ denotes the parameter space. Through optimizing objective $L^{\text{SAM}}(\theta)$, we can smooth the learning objective, thus improving the flatness of the loss landscape.

Algorithm 1 Algorithm for IGM

Input: Training set $\mathcal{D}$ and labels $\boldsymbol{Y}$;
Output: The learned parameters $\{\theta^{(j)}\}_{j=1}^{(m)}$;
INIT: Initialize gradient modification matrix. Initialize $\{\boldsymbol{T}^{(k)}\}_{j=1}^{(m)}$: $\forall k \in \{1, \cdots, m\}$, $\boldsymbol{T}^{(k)} = \boldsymbol{I}$;

```
for i = 1 → Out_Iters do
  for j = 1 → m do                                   ▷ Main iteration.
    for t = 1 → Inner_Iters do
      Randomly construct a mini-batch X_t^(j).
      Calculate loss L(θ^(j)) for data in X_t^(j).
      Calculate ε*(θ^(j)) according to Eq. (7).
      Calculate ∇_θ(j) L^SAM according to Eq. (8).
      Calculate modality index:
        k = mod(j + m − 2, m) + 1.
      Update θ^(j): θ_{t+1}^(j) = θ_t^(j) − η^(j) T^(k) ∇_θ(j) L^SAM.
    for j = 1 → n_B do                               ▷ Update {Ȳ_{n_B}^(k)}.
      Update cumulative variance according to Eq. (2).
    Update T^(j) according to Eq. (4).               ▷ Update T^(j).
```

In order to estimate the optimal perturbation ϵ^*, we can construct the following inner maximization problem [Foret *et al.*, 2021]:

$$\begin{aligned} \epsilon^*(\theta) &= \text{argmax}_{\|\epsilon\|_p \le \rho} L(\theta + \epsilon) \\ &\approx \text{argmax}_{\|\epsilon\|_p \le \rho} \epsilon^\top \nabla_\theta L(\theta) \overset{p=2}{\approx} \rho \frac{\nabla_\theta L(\theta)}{\|\nabla_\theta L(\theta)\|_2}. \end{aligned} \quad (7)$$

By substituting Equation (7) into SAM objective in Equation (6) and differentiating, we can get:

$$\begin{aligned} \nabla_\theta L^{\text{SAM}} &= \nabla_\theta \left[L\big(\theta + \epsilon^*(\theta)\big) - L(\theta) \right] + L(\theta) \\ &\approx \nabla_\theta L\left(\theta + \epsilon^*(\theta)\right) \\ &= \frac{d\left(\theta + \epsilon^*(\theta)\right)}{d\theta} \nabla_\theta L(\theta)|_{\theta + \epsilon^*(\theta)} \\ &= \nabla_\theta L(\theta)|_{\theta + \epsilon^*(\theta)} + o(\theta), \end{aligned} \quad (8)$$

where $o(\theta)$ denotes the second-order term with respect to θ and this term can be discarded to accelerate the computation. Intuitively, optimizing SAM objective can yield flatter minima which can improve the flatness of loss landscape compared with minimizing $L(\theta)$.

Since the gradient modification strategy is iterative, the SAM loss also needs to be applied to learning all modalities. Hence, the update rule in Equation (5) is modified as:

$$\boldsymbol{W}_{t+1}^{(l)} = \boldsymbol{W}_t^{(l)} - \eta^{(l)} \boldsymbol{T}^{(k)} \nabla_{\boldsymbol{W}^{(l)}} L^{\text{SAM}}(\theta^{(l)}). \quad (9)$$

The learning algorithm of IGM is summarized in Algorithm 1. In Algorithm 1, $\texttt{mod}(\cdot)$ denotes the modulo function and $\texttt{mod}(a, b)$ returns the remainder after division of a by b.

Note that the aforementioned discussion is based on the assumption that the architecture of models of different modalities is the same. In scenarios where network architectures of different modalities are heterogeneous, the gradient modification strategy can be applied to deep layers of networks with the same architecture.

4 Experiments

4.1 Datasets

We adopt five datasets, i.e., CREMA-D [Cao *et al.*, 2014], Kinetics-Sounds [Arandjelovic and Zisserman, 2017], Twitter2015 [Yu and Jiang, 2019], Sarcasm [Cai *et al.*, 2019], and NVGesture [Molchanov *et al.*, 2016], for evaluation. CREMA-D consists of 7,442 clips from 91 actors. The clips are divided into 6,698 samples for training and 744 samples for testing. Kinetics-Sounds comprises 31 human action category labels. It is divided into a training set with 15K samples, a validation set with 1.9K samples, and a testing set with 1.9K samples. Twitter2015 contains 5,338 image-text pairs with 3,179 for training, 1,122 for validation, and 1,037 for testing. Sarcasm consists of 24,635 image-text pairs. We split this dataset as 19,816 for training, 2,410 for validation, and 2,409 for testing following the setting of the original paper. NVGesture dataset contains 1,532 dynamic hand gestures. This dataset is divided into 1,050 for training and 482 for testing. We use RGB, Depth, and optical flow (OF) modalities to carry out experiments for NVGesture dataset.

4.2 Experimental Settings

Baselines: We select various methods for comparison, including OGR-GB [Wang *et al.*, 2020], OGM [Peng *et al.*, 2022], DOMFN [Yang *et al.*, 2022], MSES [Fujimori *et al.*, 2019], PMR [Fan *et al.*, 2023], AGM [Li *et al.*, 2023], MSLR [Yao and Mihalcea, 2022], ReconBoost [Hua *et al.*, 2024], sample-level modality valuation (SMV) [Wei *et al.*, 2024], MMPareto [Wei and Hu, 2024], and MLA [Zhang *et al.*, 2024]. Among these methods, OGR-GB, OGM, DOMFN, SMV, and MMPareto are early fusion methods. The remaining are late fusion methods.

Evaluation Protocols: We use accuracy (Acc.) and mean average precision (MAP) for CREMA-D and Kinetics-Sounds datasets following the setting of OGM [Yang *et al.*, 2022]. For Twitter2015, Sarcasm, and NVGesture datasets, we use accuracy and macro-F1 as evaluation metrics following the setting of the paper [Cai *et al.*, 2019]. The accuracy is used to measure the proportion of concordance between predictions and ground-truth labels. The MAP can be calculated by taking the mean of average precision for each category. And the macro-F1 can be calculated by averaging the F1 scores for each category.

Implementation Details: Following the setting of OGM, we use ResNet18 [He *et al.*, 2016] as the backbone to encode audio and video for CREMA-D and Kinetics-Sounds datasets . For Twitter2015 and Sarcasm datasets, we adopt BERT [Devlin *et al.*, 2019] as the text encoder and ResNet50 [He *et al.*, 2016] as the image encoder following the setting of the paper [Yu and Jiang, 2019]. For NVGesture dataset, we follow the data preparation steps outlined in the paper [Wu *et al.*, 2022a] and employ the I3D [Carreira and Zisserman, 2017] as unimodal branches. For a fair comparison, all baselines adopt the same backbone for the experiment. For IGM, we explore a three-layer network, which can be denoted as "FC($Dim \times 256$) $\rightarrow$ ReLU $\rightarrow$ FC(256×64) $\rightarrow$ FC($64 \times c$)", as classification head after features are extracted. Here, "FC" and "ReLU" denote the full-connected layer and ReLU [He *et al.*, 2016] layer, respectively, and "Dim" denotes the dimension of features extracted by the encoder. For audio and video modalities, the dimension of the feature is 512. For image-text modalities and NVGesture dataset, the dimension is 1024. The gradient modification strategy is applied for the classification head for IGM. Furthermore, for IGM, we use SGD as the optimizer for the audio-video and NVGesture datasets, with a momentum of 0.9 and weight decay of 1×10^{-4}. The initial learning rate is set to be 1×10^{-2}, and is divided by 10 when the loss is saturated. For image-text datasets [Yu and Jiang, 2019; Cai *et al.*, 2019], we use Adam as the optimizer, with an initial learning rate of 1×10^{-5}. By using the cross-validation strategy with a validation set, the hyper-parameter scaling factor τ is set to be 0.4 for all datasets. The hyper-parameter ρ is set to be 1×10^{-15} and 1×10^{-10} for image/text modality and audio modality, respectively. During calculating cumulative variance, we set batch size as 12 for all datasets except NVGesture. For NVGesture dataset, the batch size is set to 6 due to memory limitation. For all hyper-parameters, we utilize a cross-validation strategy on a validation set to determine their value. The experiments are performed with an NVIDIA RTX 3090 GPU.

4.3 Comparison with SOTA MML baselines

We conduct comprehensive experiments to verify the superiority of IGM. We compare IGM with SOTA MML baselines on all datasets. We report the results in Table 1, where the best and the second-best results are shown in bold and underlining, respectively. We use Unimodal-1/2/3 to denote the results based on unimodal. Unimodal-1/2 respectively denote the video/audio for CREMA-D and Kinetics-Sounds, and text/image for Twitter2015 and Sarcasm. For NVGesture dataset, Unimodal-1/2/3 denotes the RGB/OF/Depth modality, respectively. Furthermore, the results of "MLA*" are referred from the original paper of MLA. And the results of "MLA" are reproduced by us based on the authors' source code. For IGM, we adopt the same weighting strategy as the MLA method for fair comparison. We use "IGM w/o SAM" to denote IGM without SAM loss.

From Table 1, we can observe that: (1). Compared with various SOTA baselines, IGM can achieve the best performance in almost all cases by substantially large margins, including the scenarios involving two and three modalities. (2). IGM w/o SAM can outperform MLA in all cases, demonstrating that our proposed flat projection-based GM strategy achieves better performance while effectively avoiding poor plasticity. (3). IGM outperforms IGM w/o SAM in call cases, demonstrating that SAM-based optimization can further boost model performance. The underlying reasons will be discussed in ablation study section. (4). Furthermore, we find that the results of some baselines are worse than that of unimodal method, which is indicated by symbol † in Table 1.

4.4 Ablation Study

Effectiveness of GM and SAM Loss: To fully explore the effectiveness of IGM, we study the influence of different components, including the gradient modification strategy and SAM loss. The accuracy on CREMA-D dataset are

Method	CREMA-D		Kinetics-Sounds		Twitter2015		Sarcasm		NVGesture	
	Acc.	MAP	Acc.	MAP	Acc.	Mac-F1	Acc.	Mac-F1	Acc.	Mac-F1
Unimodal-1	.6317	.6861	.5312	.5669	.7367	.6849	.8136	.8065	.7822	.7833
Unimodal-2	.4583	.5879	.5462	.5837	.5863	.4333	.7181	.7073	.7863	.7865
Unimodal-3	-	-	-	-	-	-	-	-	.8154	.8183
OGR-GB	.6465	.6854†	.6710	.7139	.7435	.6869	.8335	.8271	.8299	.8305
OGM	.6694	.7173	.6606	.7144	**.7492**	.6874	.8323	.8266	-	-
DOMFN	.6734	.7372	.6625	.7244	.7445	.6857	.8356	.8262	-	-
MSES	.6156†	.6683†	.6471	.7063	.7184†	.6655†	.8418	.8360	.8112†	.8147†
PMR	.6659	.7030	.6656	.7193	.7425	.6860	.8360	.8249	-	-
AGM	.6707	.7358	.6602	.7252	.7483	.6911	.8402	.8344	.8278	.8282
MSLR	.6546	.7138	.6591	.7196	.7252†	.6439†	.8423	.8369	.8286	.8292
ReconBoost	.7484	.8124	.7085	.7424	.7442	.6834	.8437	.8317	.8413	.8632
SMV	.7872	.8417	.6900	.7426	.7428	.6817	.8418	.8368	.8352	.8341
MMPareto	.7487	.8535	.7000	<u>.7850</u>	.7358	.6729	.8348	.8284	.8382	.8424
MLA*	.7970	-	.7135	-	-	-	-	-	-	-
MLA	.7943	.8572	.7004	.7413	.7352†	.6713†	.8426	.8348	.8373	.8387
IGM w/o SAM	<u>.8026</u>	<u>.8830</u>	<u>.7159</u>	.7623	.7395	<u>.6912</u>	<u>.8455</u>	<u>.8390</u>	<u>.8487</u>	<u>.8634</u>
IGM	**.8105**	**.8948**	**.7403**	**.7855**	<u>.7489</u>	**.6917**	**.8468**	**.8392**	**.8693**	**.8703**

Table 1: Comparison with state-of-the-art multimodal learning baselines. The best and second-best performances are highlighted in **bold** and <u>underlined</u>, respectively.

SAM	GM	Audio	Video	Multi
✘	✘	45.83%	63.17%	64.52%
✔	✘	58.60%	64.79%	73.42%
✘	✔	60.13%	65.06%	80.26%
✔	✔	**61.16%**	**67.82%**	**81.05%**

Table 2: Ablation study on CREMA-D dataset.

Method	*Initial*	*Out_Iters*=1		*Out_Iters*=2	
		Audio	Video	Audio	Video
w/o a-GM	<u>.0325</u>	<u>.5312</u>	.6803	.7231	.7482
w/o v-GM	<u>.0325</u>	<u>.5312</u>	<u>.7023</u>	.7472	.7646
IGM	.0325	.5312	.7023	**.7557**	**.8105**

Table 3: Interactive enhancement analysis.

Scope of GM	Accuracy	MAP
100%	75.34%	81.23%
50%	78.97%	85.58%
30%	82.97%	90.15%
1.3% (Classification head)	81.05%	89.48%
0% (w/o GM)	73.42%	81.77%

Table 4: Results with different scope of GM.

reported in Table 2, where "SAM"/"GM" denotes whether the SAM objective/gradient modification strategy is applied during training, respectively. And "Audio", "Video", and "Multi" denote that the results based on audio, video, and multiple modalities, respectively. From Table 2, we can find that both gradient modification strategy and SAM loss can boost the performance in MML.

Necessity of Interactive Enhancement

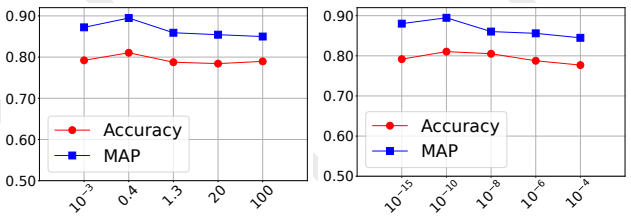
We carry out an experiment on CREMA-D dataset to further analyze the necessity of interactive enhancement. The algorithm of IGM designs an interactive learning strategy by using the gradient modification matrix of one modality to modify the gradient of another modality. To verify the effectiveness of this strategy, we design a unidirectional gradient modification experiment for comparison. Specifically, we only use the model of audio modality to modify the gradient of video modality, which is denoted as "w/o v-GM". The notation "w/o a-GM" is defined similarly. We report the results in Table 3. In Table 3, we report the accuracy after initialization in the column of "*Initial*". The other columns represent the accuracy calculated after completing the learning of a certain mode at different iterations. For w/o v-GM and w/o a-GM, the accuracy in the initial stage, the stage without applying GM strategy, and the stage with the same GM strategy is the same as IGM, which is underlined in Table 3.

From Table 3, we can observe that: (1). The performance of IGM is better than that of the unidirectional gradient modification, i.e., "w/o v-GM" and "w/o a-GM". (2). Compared with the "w/o a-GM", "w/o v-GM" can achieve better performance. In other words, the method that uses the model of the dominant modality (audio) to modify the gradient of the non-dominant modality (video) is superior to the method that uses the model of the non-dominant modality to modify the gradient of the dominant modality.

4.5 Sensitivity to Hyper-Parameters

Hyper-Parameter τ and ρ: We study the influence of hyper-parameter τ and ρ on CREMA-D dataset. We present the accuracy and MAP values with different $\tau \in [10^{-3}, 100]$ and $\rho \in [10^{-15}, 10^{-4}]$. The results are shown in Figure 3. From Figure 3, we can see that IGM is not sensitive to scaling factor τ and hyper-parameter ρ in a large range.

The Scope of Gradient Modification: In this section, we study the influence of the scope of the gradient modification.



(a). Scaling factor τ. (b). Hyper-parameter ρ.

Figure 3: Sensitivity to τ and ρ.

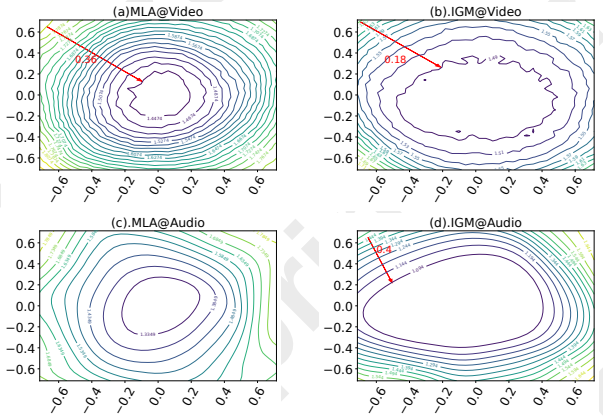


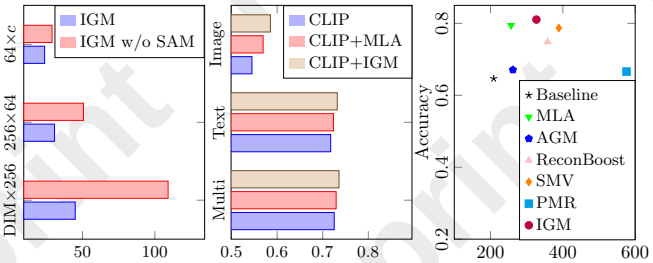
Figure 4: Loss landscape visualization.

We carry out this experiment on CREMA-D dataset, where the network architectures of audio and video modalities are the same, i.e., a ResNet18 as encoder and three full-connected layers as classification head. We select parameters along the deep to shallow layers of the neural network. And we define the scope of gradient modification as the proportion of the selected network parameters to the total parameters. The results are shown in Table 4, where "0% (w/o GM)" is used as the baseline and means that we don't perform gradient modification strategy during training. We can see that the best performance is achieved when we choose 30% parameters for gradient modification. In contrast, choosing all parameters for gradient modification does not achieve the best performance. We argue that the essence of this phenomenon is that the shallow neural network focuses on the learning of visual feature patterns, and it is not suitable for too much perturbation, especially for heterogeneous data. Furthermore, we can also find that the performance of the method applying gradient modification is better than that of the method which does not apply gradient modification.

4.6 Further Analysis

Loss Landscape Visualization: To illustrate the impact of SAM optimization, we utilize the DNN visualization method [Li *et al.*, 2018] to plot 2D loss function of MLA and IGM on CREMA-D dataset. The results of the loss landscape are shown in Figure 4. We can find that the loss change of IGM is smaller than that of MLA. That is to say, the loss landscape of our proposed method is flatter than that of MLA.

Magnitude of Singular Values: According to Equation (3),



(a). Singular values. (b). Pretrained Model. (c). Training time.

Figure 5: Analysis for singular values, robustness of the pretrained model, and training time.

the magnitude of singular values reflects the loss flatness of the direction indicated by corresponding singular vectors. We report the singular values of different layers for the IGM and the method which does not adopt SAM loss (denoted as "IGM w/o SAM"). The average singular values for different layers on CREMA-D dataset are shown in Figure 5 (a). From Figure 5 (a), we can find that the singular values of IGM are smaller than that of the IGM w/o SAM in most cases. In other words, the loss landscape of IGM is flatter than that of the method without SAM loss.

Robustness of the Pretrained Model: We further explore the robustness of the large vision-language pre-trained model on Twitter2015 dataset. Following the setting of MLA [Zhang *et al.*, 2024], we replace the backbones of image and text modalities as the corresponding encoders of CLIP [Radford *et al.*, 2021]. We adopt the same three-layer network as the classification head for multimodal learning. Then we fine-tune the model on Twitter2015 dataset. We report accuracy results in Figure 5 (b), where "CLIP+MLA" and "CLIP+IGM" denote that during fine-tuning we apply MLA and IGM, respectively. From Figure 5 (b), we can find that: (1). MLA and IGM can achieve better performance compared with CLIP. (2). IGM can boost higher improvement based on CLIP encoder compared with MLA.

Training Overhead: We compare the training overhead of IGM with competitive state-of-the-art baselines, including Baseline, AGM, PMR, MLA, and ReconBoost, through empirical experiments under the same setting on CREMA-D dataset. The results are shown in Figure 5 (c), where the training times are reported in hours. It can be observed that IGM achieves the best accuracy while maintaining competitive training time.

5 Conclusion

In this paper, we propose a novel MML method, called interactive MML via flat gradient modification (IGM). We first employ a flat projection-based gradient modification strategy to enhance the interaction during learning and avoid poor plasticity issue. Furthermore, we introduce SAM-based optimization to fully exploit the flatness of the learning objective, further smoothing the learning objective. To this end, IGM can further mitigate the modality imbalance problem and lead to better performance. Extensive experiments demonstrate the superiority of IGM compared with various SOTA methods across five widely used datasets.

Acknowledgments

This work is supported by the National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

References

[Arandjelovic and Zisserman, 2017] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617. IEEE, 2017.

[Cai *et al.*, 2019] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pages 2506–2515. Association for Computational Linguistics, 2019.

[Cao *et al.*, 2014] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: crowd-sourced emotional multimodal actors dataset. *TAC*, 5(4):377–390, 2014.

[Carreira and Zisserman, 2017] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE, 2017.

[Chang *et al.*, 2015] Angel X. Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. Text to 3d scene generation with rich lexical grounding. In *ACL*, pages 53–62. The Association for Computer Linguistics, 2015.

[Chaudhari *et al.*, 2017] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *ICLR*. OpenReview.net, 2017.

[Deng *et al.*, 2021] Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. In *NeurIPS*, pages 18710–18721, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Du *et al.*, 2022] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Yue Wang, Yang Yuan, and Hang Zhao. Modality laziness: Everybody's business is nobody's business. In *ICLR*. OpenReview.net, 2022.

[Du *et al.*, 2023] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *ICML*, volume 202, pages 8632–8656. PMLR, 2023.

[Fan *et al.*, 2023] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: prototypical modal rebalance for multimodal learning. In *CVPR*, pages 20029–20038. IEEE, 2023.

[Foret *et al.*, 2021] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*. OpenReview.net, 2021.

[Fujimori *et al.*, 2019] Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-specific learning rate control for multimodal classification. In *ACPR*, volume 12047, pages 412–422, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.

[Hua *et al.*, 2024] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *ICML*. PMLR, 2024.

[Huang *et al.*, 2023] Zhuo Huang, Li Shen, Jun Yu, Bo Han, and Tongliang Liu. Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning. In *NeurIPS*, 2023.

[Keskar *et al.*, 2017] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*. OpenReview.net, 2017.

[Li *et al.*, 2018] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, pages 6391–6401, 2018.

[Li *et al.*, 2020] Jing Li, Jing Xu, Fangwei Zhong, Xiangyu Kong, Yu Qiao, and Yizhou Wang. Pose-assisted multi-camera collaboration for active object tracking. In *AAAI*, pages 759–766. AAAI Press, 2020.

[Li *et al.*, 2023] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pages 22157–22167. IEEE, 2023.

[Liang *et al.*, 2022] Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. AF: an association-based fusion method for multi-modal classification. *TPAMI*, 44(12):9236–9254, 2022.

[Liu *et al.*, 2024] Bo Liu, Lejian He, Yuchen Xie, Yuejia Xiang, Li Zhu, and Weiping Ding. Minjot: Multimodal infusion joint training for noise learning in text and multimodal classification problems. *INFFUS*, 102:102071, 2024.

[Molchanov *et al.*, 2016] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *CVPR*, pages 4207–4215. IEEE, 2016.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696. Omnipress, 2011.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal

learning via on-the-fly gradient modulation. In *CVPR*, pages 8228–8237. IEEE, 2022.

[Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951. AAAI Press, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763. PMLR, 2021.

[Sun *et al.*, 2022] Qing Sun, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Exploring example influence in continual learning. In *NeurIPS*, 2022.

[Wang *et al.*, 2016] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *CoRR*, abs/1607.06215, 2016.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12692–12702. IEEE, 2020.

[Wang *et al.*, 2021] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *CVPR*, pages 184–193. IEEE, 2021.

[Wei and Hu, 2024] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*. PMLR, 2024.

[Wei *et al.*, 2024] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pages 27338–27347. IEEE, 2024.

[Wu *et al.*, 2022a] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, volume 162, pages 24043–24055. PMLR, 2022.

[Wu *et al.*, 2022b] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pages 24043–24055. PMLR, 2022.

[Xiao *et al.*, 2022] Fangxiong Xiao, Lixi Deng, Jingjing Chen, Houye Ji, Xiaorui Yang, Zhuoye Ding, and Bo Long. From abstract to details: A generative multimodal fusion framework for recommendation. In *ACMMM*, pages 258–267. ACM, 2022.

[Yang *et al.*, 2019] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *IJCAI*, pages 4092–4098. ijcai.org, 2019.

[Yang *et al.*, 2022] Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu. DOMFN: A divergence-orientated multi-modal fusion network for resume assessment. In *ACMMM*, pages 1612–1620. ACM, 2022.

[Yang *et al.*, 2024a] Yang Yang, Jinyi Guo, Guangyu Li, Lanyu Li, Wenjie Li, and Jian Yang. Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning. *FCS*, 18(3):181335, 2024.

[Yang *et al.*, 2024b] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, 2024.

[Yang *et al.*, 2025] Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *TPAMI*, 2025.

[Yao and Mihalcea, 2022] Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *ACL*, pages 1824–1834. Association for Computational Linguistics, 2022.

[Ye *et al.*, 2018] Yongkai Ye, Xinwang Liu, Qiang Liu, Xifeng Guo, and Jianping Yin. Incomplete multiview clustering via late fusion. *CIN*, 2018:6148456:1–6148456:11, 2018.

[Yu and Jiang, 2019] Jianfei Yu and Jing Jiang. Adapting BERT for target-oriented multimodal sentiment classification. In *IJCAI*, pages 5408–5414. ijcai.org, 2019.

[Zhang *et al.*, 2024] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, pages 27456–27466. IEEE, 2024.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398. IJCAI/AAAI Press, 2016.

[Zhao *et al.*, 2023] Zhen Zhao, Zhizhong Zhang, Xin Tan, Jun Liu, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Rethinking gradient projection continual learning: Stability/plasticity feature space decoupling. In *CVPR*, pages 3718–3727. IEEE, 2023.

[Zhu *et al.*, 2023] Lei Zhu, Tianshi Wang, Fengling Li, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *CoRR*, abs/2308.14263, 2023.

[Zhu *et al.*, 2024] Linlin Zhu, Heli Sun, Qunshu Gao, Tingzhou Yi, and Liang He. Joint multimodal aspect sentiment analysis with aspect enhancement and syntactic adaptive learning. In *IJCAI*, pages 6678–6686. ijcai.org, 2024.

[Zong *et al.*, 2024] Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, and Ken Zheng. Balancing multimodal learning via online logit modulation. In *IJCAI*, pages 5753–5761. ijcai.org, 2024.