Improving Multimodal Learning via Imbalanced Learning

Shicai Wei[1,2] Chunbo Luo[1]* Yang Luo[1]
[1]University of Electronic Science and Technology of China
[2] National and Local Joint Engineering Research Center for Cloud Operating System
shicaiwei@std.uestc.edu.cn, {c.luo, luoyang}@uestc.edu.cn

Abstract

Multimodal learning often encounters the under-optimized problem and may perform worse than unimodal learning. Existing approaches attribute this issue to imbalanced learning across modalities and tend to address it through gradient balancing. However, this paper argues that balanced learning is not the optimal setting for multimodal learning. With bias-variance analysis, we prove that imbalanced dependency on each modality obeying the inverse ratio of their variances contributes to optimal performance. To this end, we propose the Asymmetric Representation Learning(ARL) strategy to assist multimodal learning via imbalanced optimization. ARL introduces auxiliary regularizers for each modality encoder to calculate their prediction variance. ARL then calculates coefficients via the unimodal variance to re-weight the optimization of each modality, forcing the modality dependence ratio to be inversely proportional to the modality variance ratio. Moreover, to minimize the generalization error, ARL further introduces the prediction bias of each modality and jointly optimizes them with multimodal loss. Notably, all auxiliary regularizers share parameters with the multimodal model and rely only on the modality representation. Thus the proposed ARL strategy introduces no extra parameters and is independent of the structures and fusion methods of the multimodal model. Finally, extensive experiments on various datasets validate the effectiveness and versatility of ARL. Code is available at https://github.com/shicaiwei123/ICCV2025-ARL

1. Introduction

Multimodal learning has gained considerable attention across a wide range of applications, including action recognition [2, 31], object detection [16, 26], and audiovisual speech recognition [1, 24]. While multimodal learning has significant potentiality in improving the robustness and performance of models, the heterogeneity of multimodal data raises the challenges to leverage multimodal correlations and complementarities. According to recent research [9, 10, 21], the performance of existing multimodal methods is sub-optimal since the performance of each modality encoder falls significantly short of their unimodal bound.

Existing methods attribute this under-optimized phenomenon to the "modality imbalance" problem, where the dominant modality suppresses the learning of weak modality, hindering the full utilization of multimodal data [9, 10, 18, 21, 28, 40]. They then propose several methods to address the problem. Some of them aid the training of the weak modality with the help of additional unimodal classifiers [28, 34], pre-trained models [9] or modality resampling [35]. Since they will introduce additional neural modules and complicate the training procedure, other methods [10, 18, 21] try to modulate the gradient of different modalities within the multimodal model. Moreover, recent methods transform the conventional joint multimodal learning process into an alternating unimodal learning process to minimize inter-modality interference directly [15, 36, 40].

While existing methods achieve good performance, they simply attribute the inferior multimodal performance to the imbalanced learning between modalities. This paper argues that this assertion may be imperfect. According to the assertion, for the CREMA-D datasets where the audio branch outperforms the visual branch (see Fig. 1 (a)), increasing the gradient of the audio branch should exacerbate the imbalanced learning and decrease performance. However, when we increase the gradient of the audio branch by 5 times, the model performance does not decrease but increases (see Fig. 1 (b)). This finding drives us to further explore the real bottleneck that limits multimodal learning performance.

To this end, we reframe the conventional joint multimodal learning process by transforming it into the ensembling learning of multiple unimodal processes. From the perspective of bias-variance, we prove that the optimal decision dependence ratio on each modality should obey the inverse of their prediction variance. And conventional balanced learning is optimal only when the variances of the different modalities are the same. Therefore, instead of balancing multimodal learning, we should modulate the gradi-

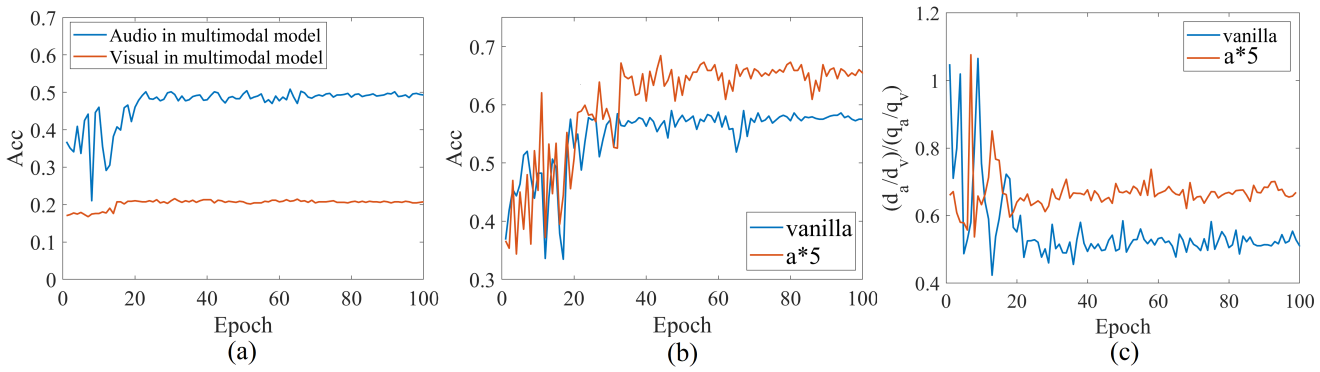

Figure 1. The visualization on the audio-visual dataset CREMA-D. (a) presents the performance of the unimodal branch in the vanilla multimodal model. (b) presents the performance of the multimodal model with audio branch modulation. '$a * 5$' represents increasing the gradient for the audio branch by fivefolds to accelerate its optimization. (c) presents the gap between the modality dependence and modality variance. $\frac{d_a}{d_v}$ denotes the model optimization dependence ratio on the audio and visual modalities (see Equation 4). $\frac{q_a}{q_v}$ denotes the inverse of the variance ratio of audio and visual modalities (see Equation 17).

ent of each modality to adjust their contribution ratio to the model optimization, enabling it to satisfy the inverse of the modality variance ratio. As shown in Fig. 1 (c), increasing the gradient of the audio branch in the multimodal model exactly alleviates the gap between the optimization dependence and the inverse of modality variance ratio.

To this end, we propose the Asymmetric Representation Learning (ARL) strategy to assist multimodal learning. ARL first introduces auxiliary regularizers for each modality encoder, which measure their variance based on the prediction logit output. Then ARL introduces asymmetric coefficients to modulate the gradients of each modality to adjust the optimization dependency on each modality, aligning it with the inverse of the modality variance ratio. Additionally, ARL incorporates the original gradient as a residual component to the adjusted gradient, ensuring that encoder optimization continues even when the modulation factor is extremely small. Furthermore, to reduce generalization error, ARL integrates the prediction bias of each modality and optimizes them in conjunction with the multimodal loss. Importantly, all auxiliary regularizers share parameters with the multimodal model and depend solely on the modality representations. Thus, the ARL strategy introduces no extra parameters and is agnostic to the architecture and fusion techniques of the multimodal model. We conduct extensive experiments in various multimodal tasks on different datasets to study the effectiveness and versatility of ARL. The contributions are summarized as follows,

- We reveal that the balanced learning of different modalities is not the optimal multimodal learning setting. On the contrary, we prove that imbalanced optimization obeying the inverse of modality variance ratio can contribute to optimal performance. This provides an insightful view to study under-optimized multimodal learning.
- We propose the ARL strategy to modulate the gradients of each modality, accelerating their optimization and aligning their contribution to model optimization with their variance.
- Extensive experiments on various tasks and datasets demonstrate the effectiveness and generalization of the proposed ARL in boosting multimodal learning.

2. Related Works

2.1. Multimodal Learning

Models fusing data from multiple modalities have shown superior performance over unimodal models in various applications, such as action recognition [2, 4], object detection [16, 26, 32], and audiovisual speech recognition [1, 24]. Researchers have pursued various research directions based on specific applications. For instance, some have delved into investigating the fusion strategy and network framework to improve the model's robustness to incomplete multimodal data [8, 30, 33]. Furthermore, substantial efforts have been dedicated to harnessing information from multiple modalities to boost model performance in specialized tasks. These tasks encompass action recognition [7, 11, 19], semantic segmentation [6, 14, 22] and audio-visual speech recognition [1, 24]. Nonetheless, most multimodal methods employing joint training strategies could not fully exploit all modalities and produce under-optimized unimodal representations. Consequently, the performance of multimodal models falls short of expectations, and even cannot achieve better performance compared to the best single-modal DNNs [28].

2.2. Under-optimized Multimodal Learning

The aforementioned defect of multimodal learning methods encourages researchers to explore the reasons behind it. Wang *et al*. [28] found that different modalities overfit and generalize at different rates and thus obtain suboptimal solutions. To this end, they calculate the logit output of each modality and their fusion and then optimize the gradient mixing problem to obtain better weights for each branch.

Furthermore, some works [9, 10, 18, 21, 35, 37, 40] proposed that the better-performing modality will dominate the gradient update while suppressing the learning process of the other modality. To this end, Du *et al*. [9] aimed to enhance unimodal performance by distilling knowledge from well-trained models. However, this introduces more model structure and computational effort, making the training process more complex and expensive. Therefore, OGM[21] chose to reduce the gradient of the dominant modality to mitigate its inhibitory effect on the other modalities. Xu *et al*. [37] then extended OGM to the multi-modal fine-grained task via the modality-wise L2 normalization to features and weights. Although a certain degree of improvement is achieved, such approaches do not impose the intrinsic motivation of improvement on the slow-learning modality. Thus, Fan *et al*. [10] introduced the prototypical rebalance strategy to accelerate the slow-learning modality and alleviate the suppression of the dominant modality. Furthermore, MLA [40] and ReconBoost [15] transformed the vanilla joint multimodal learning process into an alternating unimodal process to minimize inter-modality interference directly. MMPareto [34] introduced Pareto integration to alleviate the gradient conflict between multimodal and unimodal learning objectives. D&R [36] estimated the separability of each modality in its unimodal representation space, and then used it to softly re-initialize the corresponding unimodal encoder

Generally, these methods aim to improve multimodal learning via unimodal assistance or balanced learning. This considers all modalities to be equally important, ignoring their inherent capacity discrepancy. This paper argues that balanced optimization is not the optimal setting for multimodal learning. On the contrary, imbalanced learning obeying modality variance can contribute to better performance.

3. Method

We will first re-analyze the under-optimized problem in multimodal learning and prove that a multimodal model can achieve optimal performance when the optimization dependencies on each modality are consistent with the inverse of their variance ratio. Then we introduce the ARL strategy to boost multimodal learning according to this theory.

3.1. Re-analyze the Under-optimized Problem in Multimodal Learning.

We will first introduce the basic model for multimodal learning. Then we re-analyze the under-optimized phenomenon in multimodal learning.

Multimodal learning model. Without loss of generality, we consider two input modalities as m_0 and m_1. The dataset is denoted as $\mathcal{D} = \{x_i^{m_0}, x_i^{m_1}, y_i\}_{i=1,2,\ldots,N}$, where $y \in \{1, 2, \ldots, M\}$, and M is the number of categories. We use two encoders $\varphi_0(\boldsymbol{\theta}_0, \cdot)$ and $\varphi_1(\boldsymbol{\theta}_1, \cdot)$ to extract features, where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta_1}$ are the parameters of encoders. The representation outputs of encoders are denoted as $\boldsymbol{z}_0 = \phi_0(\boldsymbol{\theta}_0, {x_i}^{m_0})$ and $\boldsymbol{z}_1 = \phi_1(\boldsymbol{\theta}_1, {x_i}^{m_1})$. The two unimodal encoders are connected through the representations by some kind of fusion methods, which is prevalent in multimodal learning [7, 13, 29]. Here, let $\phi_f(\boldsymbol{\theta_f}, .)$ denotes the fusion module. $\boldsymbol{\theta}_f$ is the parameter of this module. Let $\boldsymbol{W} \in \mathbb{R}^{M\times(d_0+d_1)}$ and $\boldsymbol{b} \in \mathbb{R}^M$ denote the parameters of the linear classifier to produce the logits output. The output of input x_i in a multimodal model can be expressed as follows,

$$\begin{cases} f(x_i) = \boldsymbol{W}\boldsymbol{z}_f + \boldsymbol{b} & (1) \\ \boldsymbol{z}_f = \phi_f(\boldsymbol{\theta_f}, \boldsymbol{z}_0; \boldsymbol{z}_1) & (2) \end{cases}$$

Here, take the most widely used vanilla fusion method, concatenation, as an example, $\boldsymbol{z}_f = [\boldsymbol{z}_0; \boldsymbol{z}_1]$ and thus $f(x_i)$ can be rewritten as follows,

$$f(x_i) = \boldsymbol{W}_0 \cdot \boldsymbol{z}_0 + \boldsymbol{b}_0 + \boldsymbol{W}_1 \cdot \boldsymbol{z}_1 + \boldsymbol{b}_1 \tag{3}$$

where $\boldsymbol{W} = [\boldsymbol{W}_0; \boldsymbol{W}_1]$, $\boldsymbol{W}_0 \in \mathbb{R}^{M\times d_0}, \boldsymbol{W}_1 \in \mathbb{R}^{M\times d_1}$ and $\boldsymbol{b}_0, \boldsymbol{b}_1 \in \mathbb{R}^M$, respectively.

Let $\boldsymbol{s}_i^{m_0} = \boldsymbol{W}_0 \cdot \boldsymbol{z}_0 + \boldsymbol{b}_0$, which denotes the logit output of modality m_0, and $\boldsymbol{s}_i^{m_1} = \boldsymbol{W}_1 \cdot \boldsymbol{z}_1 + \boldsymbol{b}_1$, which denotes the logit output of modality m_1. As a result, the final output is the summation of $\boldsymbol{s}_i^{m_0}$ and $\boldsymbol{s}_i^{m_1}$. Thus, the gradient is determined by $\boldsymbol{s}_i^{m_0}$ and $\boldsymbol{s}_i^{m_1}$ and the optimization dependency coefficient d is defined as follows,

$$d = \frac{d^{m_0}}{d^{m_1}} = \frac{exp(\boldsymbol{s}_i^{m_0}(y_i))/\sum_{j=1}^{M} exp(\boldsymbol{s}_i^{m_0}(j))}{exp(\boldsymbol{s}_i^{m_1}(y_i))/\sum_{j=1}^{M} exp(\boldsymbol{s}_i^{m_1}(j))} \tag{4}$$

Existing studies [10, 18, 21] hold that the modality with better performance will dominate the optimization of the weak one and lead to insufficient multimodal learning. Thus they try to rebalance the learning of different modalities by gradient modulation to make d equal 1.

However, as shown in Fig. 1 (b), when we increase the gradient of the perform-better modality by 5 times, enlarging the training imbalance, the model performance does not decrease but increases. This finding drives us to further explore the real bottleneck that limits multimodal learning

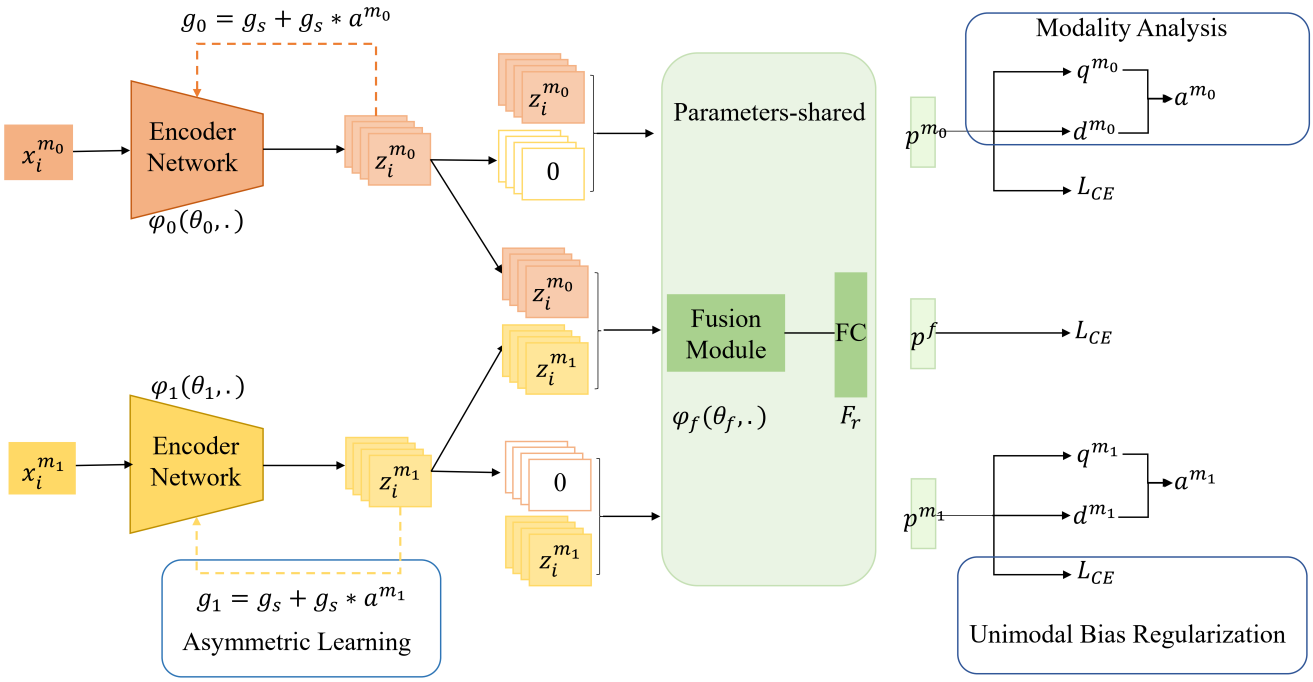

Figure 2. The pipeline of the multimodal model with ARL strategy. It consists of three components: the modality analysis to measure the modality variance; the asymmetric learning to align the optimization dependency on each modality with the inverse of their variance ratio; and the unimodal bias regularization to reduce multimodal bias. g_s is the gradient passed back from the fusion module.

performance. More specifically, *what is the optimal optimization dependency ratio for different modalities?*

The optimal optimization dependency ratio for different modalities. As discussed above, balancing the contribution of different modalities to multimodal model optimization is not optimal. This drives us to find the optimal contribution ratio of different modalities. Generally, the optimal contribution ratio should minimize the generalization error, thus we formulate the problem as follows,

$$\begin{cases} \min\limits_{w_0,w_1} g = \dfrac{1}{N}\sum\limits_{i=1}^{i=N} L(f(x_i), y_i) & (5) \\ f(x_i) = w_0 s_i^{m_0} + w_1 s_i^{m_1} & (6) \end{cases}$$

where $w_0 > 0$ and $w_1 > 0$ denotes the contribution of modality m_0 and m_1, respectively, and $w_0 + w_1 = 1$. $L(.)$ measures the generalization error from prediction and groundtruth. For simplification, we define $E[f(x)] = \frac{1}{N}\sum_{i=1}^{i=N} f(x_i)$.According to the bias-variance decomposition [23], the error between prediction and groundtruth can be rewritten as follows,

$$\begin{cases} g = (Bias(f(x), y))^2 + Var(f(x)) + Var(\epsilon) & (7) \\ Bias(f(x), y) = E[f(x) - y] & (8) \\ Var(f(x)) = E\left[f(x)^2\right] - E[f(x)]^2 & (9) \end{cases}$$

where $Bias(f(x), y)$ is the bias of prediction, which measures the error of the prediction. $Var(f(x))$ is the variance of $f(x)$, which measures the uncertainty of the prediction, $Var(\epsilon)$ is the irreducible error in the dataset and cannot be reduced by any model. In other words, to minimize the Eq. 5, we need to minimize the $Bias(f(x), y)^2$ and $Var(f(x))$. Limited by page, we only give the key results here and detailed derivation can be seen in the supplementary materials.

For the term of $Bias(.)^2$, we can convert it as follows,

$$Bias(f(x), y)^2 = (w_0 Bias(s^{m_0}, y) + w_1 Bias((s^{m_1}, y))^2 \quad (10)$$

Then, with the constraint $w_0 + w_1 = 1$, we get the solution of w_0 and w_1 as follows,

$$\begin{cases} w_0 = \dfrac{Bias(s^{m_1}, y)}{Bias(s^{m_1}, y) - Bias(s^{m_0}, y)} & (11) \\ w_1 = \dfrac{-Bias(s^{m_0}, y)}{Bias(s^{m_1}, y) - Bias(s^{m_0}, y)} & (12) \end{cases}$$

Here, the numerical solution is meaningless since one of w_0 or w_1 must be smaller than 0, conflicting with $w_1 > 0$ and $w_1 > 0$. Thus, when $Bias(s^{m_0}, y)$ and $Bias(s^{m_1}, y)$ are fixed, we can not find a reasonable combination of w_0

or w_1 to minimize $Bias(.)^2$. Consequently, the only way to minimize $Bias(.)^2$ is to minimize $Bias(s^{m_0}, y)$ and $Bias(s^{m_1}, y)$.

For the term of $Var(f(x))$, we can convert it as follows,

$$Var(f(x)) = {w_0}^2 Var(s^{m_0}) + {w_1}^2 Var(s^{m_1}) \tag{13}$$

Then, with the constraint $w_0 + w_1 = 1$, we get the solution of w_0 and w_1 as follows,

$$\begin{cases} w_0 = \dfrac{Var(s^{m_1})}{Var(s^{m_1}) + Var(s^{m_1})} & (14) \\ w_1 = \dfrac{Var(s^{m_0})}{Var(s^{m_1}) + Var(s^{m_1})} & (15) \end{cases}$$

. Here,we can get the relationship between $\frac{w_0}{w_1}$ and $Var(s^{m_0})$ as well as $Var(s^{m_1})$ as follows,

$$\frac{w_0}{w_1} = \frac{\frac{1}{Var(s^{m_0})}}{\frac{1}{Var(s^{m_1})}} \tag{16}$$

. In other words, to minimize the $Var(f(x))$ of multimodal models, we should enable the modality with a lower variance to have a larger contribution to model optimization.

3.2. Asymmetric Representation Learning

As discussed, balanced learning is not optimal. In contrast, imbalanced learning inversely proportional the modality variance ratio can contribute to better performance. Moreover, because there is no reasonable combination to minimize the fusion bias, we need to minimize the unimodal bias to achieve a lower generalization error. To this end, we propose a simple but effective Asymmetric Representation Learning (ARL) strategy to improve the performance of multimodal learning. As shown in Fig 2, it consists of three components: modality analysis, asymmetric learning, and unimodal bias regularization.

Modality Analysis This stage determines the inverse of the variance q of each modality. According to the definition, we need to calculate their logit output first. To this end, we introduce auxiliary regularizers for each modality to calculate their logit output directly. Compared to using s^{m_0} and s^{m_1} calculated in vanilla concatenation fusion, this is independent of model structures and fusion methods, making it highly versatile for complex scenarios. Notably, As shown in Fig 2, this introduces no extra parameters since all regularizers share the parameters with the multimodal model and calculate the unimodal loss by setting other modality representations to zero.

Since the vanilla bias-variance decomposition is designed for the regression task, we need to make it suitable for the classification task. Following existing work [12, 39], we adapt the self-information entropy to approximate the variance term in the classification task. Take the modality m_0 as an example, we measure q^{m_0} as follows,

$$q^{m_0} = 1/\sum_{j=1}^{j=M} \sigma(p^{m_0})(j) log \sigma(p^{m_0})(j) \tag{17}$$

where p^{m_0} and p^{m_1} is the logit output of modality m_0 and m_1, respectively. $\sigma(.)$ is the softmax function.

Asymmetric Learning This stage calculates the modulation coefficient a^{m_0} and a^{m_1} for each modality encoder to adjust their modality contribution ratio so that it can be inversely proportional to the modality variance ratio.

To make the optimization dependency ratio $\frac{d^{m_0}}{d^{m_1}}$ equal the modality variance ratio $\frac{q^{m_0}}{q^{m_1}}$, we should increase the gradient of m_0 to increase $\frac{d^{m_0}}{d^{m_1}}$ when it is smaller than $\frac{q^{m_0}}{q^{m_1}}$. Or we should increase the gradient of m_1 to decrease $\frac{d^{m_0}}{d^{m_1}}$ when it is larger than $\frac{q^{m_0}}{q^{m_1}}$. Therefore, we define the modulation coefficient a^{m_0} and a^{m_1} as follows,

$$[a^{m_0}, a^{m_1}] = \sigma(\left[\frac{q^{m_0}}{q^{m_1}} * T, \frac{d^{m_0}}{d^{m_1}} * T\right]) \tag{18}$$

where $\frac{d^{m_0}}{d^{m_1}}$ are calculated by Eq. 4 with $s^{m_0} = p^{m_0}$ and $s^{m_1} = p^{m_1}$. This represents the real modality dependency ratio in the current stage. $\sigma(.)$ is the softmax function. T is the temperature coefficient, which controls the intensity of the modulation. Larger T will increase the gap between a^{m_0} and a^{m_1}, accelerating the modulation. Thus the gradient g_s from shared modules passed back to modality encoder m_0 and m_1 are regulated as follows,

$$\begin{cases} g_s^{m_0} = g_s + g_s * a^{m_0} & (19) \\ g_s^{m_1} = g_s + g_s * a^{m_1} & (20) \end{cases}$$

where the residual sum with the original gradient is employed to prevent the parameter updating from stopping when the modulation factor is extremely small.

Unimodal Bias Regularization Since there is no reasonable combination that minimizes the fusion bias, this stage ARL integrates the prediction bias of each modality and optimizes them in conjunction with the multimodal loss to further reduce the generalization error.

Take the modality m_0 as an example, we measure its bias u^{m_0} from its logit output as follows,

$$u^{m_0} = L_{CE}(p^{m_0}, y) \tag{21}$$

where L_{CE} is the cross entropy loss.

Total loss. As discussed above, the total loss L_{ARL} for multimodal learning with ARL is defined as follows,

$$L_{ARL} = L_{CE}(p^f, y) + \gamma(u^{m_0} + u^{m_1}) \tag{22}$$

where p^f is the multimodal logit output. $L_{CE}(p^f, y)$ optimize the prediction error of the multimodal model. $u^{m_0} + u^{m_1}$ optimize the unimodal bias of the multimodal model.

The training pseudocode of the multimodal model with ARL is given below,

Algorithm 1 Multimodal learning with ARL strategy

Input: Training dataset D, iteration number T.
for t = 0, · · · , T **do**
  Feed-forward the batched data to the model.
  Calculate unimodal and multimodal loss via Eq.22.
  Calculate backward gradient.
  Calculate the variance of each modality via Eq.17.
  Calculate the modulation coefficient via Eq.18.
  Modify the gradient of each modality via Eq.19.
  Update model parameters via the modified gradient.

3.3. Comparison with Existing Method

Existing balance-based methods [10, 18, 21, 34] try to modulate the learning of different modalities to balance their optimization. However, we found that imbalanced learning obeying the inverse of the modality variance ratio can contribute to better performance. Therefore, we proposed ARL technology to modulate the learning of different modalities, forcing their contribution to model optimization to satisfy the inverse of their variance ratio.

4. Experiments

4.1. Datasets

CREMA-D [5] is an audio-visual dataset for emotion recognition, which consists of audio and visual modalities. There are a total of 7442 videos in the dataset for 6 usual emotions. They are further divided into 6698 samples as the training set and 744 samples as the testing set.

AVE [27] is an audio-visual video dataset for audio-visual event localization. It comprises 4,143 10-second videos across 28 event classes. Here, we extract frames from event-localized video segments and capture corresponding audio clips, creating a labeled multimodal classification dataset. The training and validation split of the dataset follows [27].

Kinetics-Sounds (KS) [3] is a dataset derived from the Kinetics dataset, focusing on 34 human action classes. Each class is chosen to be potentially manifested visually and aurally. This dataset consists of 19k 10-second video clips, with a distribution of 15k for training, 1.9k for validation, and 1.9k for testing.

MOSI [38] is a well-known dataset in the field of multimodal sentiment analysis, encompassing audio, visual, and textual data. It comprises 2,199 video clips of individual utterances, extracted from 93 monologue videos. Each clip is annotated with a sentiment score on a continuous scale from -3 (strongly negative sentiment) to 3 (strongly positive sentiment). In this paper, we use the two-class label setting that considers positive/negative results only. We employ this dataset to demonstrate the method's capability to generalize in environments involving multiple modalities.

UCF-101 [25] is a multimodal dataset for human action recognition, offering both RGB and optical flow sequences. This dataset comprises 101 distinct action categories. According to its original protocol, 9,537 video samples are used for training and 3,783 for testing.

4.2. Experimental settings

Implementation details. To ensure a fair comparison, we utilize ResNet18 as the encoder backbone for all datasets, as is common in previous methods [10, 21]. For CREMA-D, we select one frame from each clip and resize it to 224x244 as the visual input. The audio data is converted into a spectrogram of size 257×299 using librosa [20]. For AVE and Kinetics-Sounds datasets, we uniformly sample 3 frames from each video clip and resize them to 224x224 as visual inputs. The entire audio data is transformed into a spectrogram of size 257×1,004. The training configurations mirror those of previous methods [10, 21], including a mini-batch size of 64, an SGD optimizer with momentum 0.9, a learning rate of 1e-3, and a weight decay of 1e-4. γ is set as 4 for all datasets. T is set 8,4,4,4,4 for CREMA-D, AVE, KS, MOSI, and UCF101 datasets, respectively. More ablation studies are provided in the supplementary material.

Comparison settings. To study the advantage of ARL, we make comparisons with three gradient modulation approaches, OGM [21], AGM [18], PMR [10], and four unimodal regularization methods, G-Blending [28], MLA [40], MMPareto [34] and D&R [36]. For a fair comparison, we unify the backbone as ResNet18 and the fusion method as concatenation in all experiments. Note that the original MLA uses epoch as 150 and batch size as 16. We unify them with other comparison methods as epoch 100 and batch size 64. Besides, the learning rate of vanilla MMPareto [34] and D&R [36] is 2e-3, we also unify it as 1e-3.

4.3. Comparison on multimodal tasks

Comparison with other modulation strategies. The results are shown in Table 1. Firstly, the proposed ARL strategy achieves the best performance on all datasets with a significant improvement compared to existing methods. Compared to the second-best method, it improves the performance by 3.02%, 2.28%, 3.01%, and 0.98% on CREMA-D, AVE, KS, respectively. These results demonstrate the effectiveness of the ARL strategy.

More importantly, while existing balance-based methods, such as OGM-GE, AGM, and PMR, can improve the performance on the CREMA-D, KS, and MOSI datasets, they decrease the performance on the AVE dataset. This is because the balance-based methods could hinder the learn-

Dataset	CREMA-D		KS		AVE	
Methods	Acc	macro F1	Acc	macro F1	Acc	macro F1
Audio-only	57.27	57.89	48.67	48.89	62.16	58.54
Visual-only	62.17	62.78	52.36	52.67	31.40	29.87
Concatenation	58.83	59.43	64.97	65.21	66.15	62.46
Grad-Blending	68.81	69.34	67.31	67.68	67.40	63.87
OGM-GE	64.34	64.93	66.35	66.76	65.62	62.97
AGM	67.21	68.04	65.61	65.99	64.50	61.49
PMR	65.12	65.91	65.01	65.13	63.62	60.36
MMPareto	70.19	70.82	69.13	69.05	68.22	64.54
MLA	73.21	73.77	<u>69.62</u>	<u>69.98</u>	<u>70.92</u>	<u>67.23</u>
D&R	<u>73.52</u>	<u>73.96</u>	69.10	69.36	69.62	64.93
ARL	**76.61**	**77.14**	**74.28**	**74.03**	**72.89**	**68.04**

Table 1. Comparison with existing modulation strategies on CREMA-D, Kinetics-Sounds, and AVE datasets. The proposed ARL achieves the best performance. Bold and underline mean the best and second-best results, respectively.

Dataset	MOSI		UCF101	
Methods	Acc	macro F1	Acc	macro F1
Concatenation	76.92	75.68	80.41	79.40
Grad-Blending	78.16	76.94	81.73	80.84
OGM-GE	77.33	76.23	81.15	80.36
AGM	77.65	76.49	81.55	80.36
PMR	77.48	76.34	81.36	80.37
MMPareto	78.17	76.92	81.98	80.64
MLA	78.87	77.21	82.01	<u>81.22</u>
D&R	<u>78.96</u>	<u>77.37</u>	<u>82.11</u>	80.87
ARL	**79.94**	**78.84**	**83.22**	**81.98**

Table 2. **(Left)**: Comparison with imbalanced multimodal learning methods on MOSI dataset with three modalities.**(Right)**: Comparison with imbalanced multimodal learning methods on the UCF101 dataset with 101 categories. Bold and underline mean the best and second-best results, respectively.

ing of performance-dominant modality (audio modality) when they balance the optimization of audio and visual modalities. In contrast, the proposed ARL also improves the performance on the AVE dataset by 5.75% compared to the concatenation baseline. This confirms the superiority to adjust the optimization dependency on each modality obeying their variance ratio.

Comparison in more-than-two modality case. Existing methods primarily focus on handling datasets containing two modalities [10, 18, 21], which has certain limitations when dealing with broader scenarios. In contrast, the ARL method is not constrained by the number of modalities and demonstrates strong adaptability. In this study, we conducted experimental validation on the MOSI dataset, which includes audio, visual, and textual modalities. For a comprehensive comparison, we extend the core unimodal balancing strategies of G-Blending, OGM-GE, AGM, PMR, and MMPareto to handle cases involving more than two modalities. As shown in the left part of Table 2, ARL also achieves the best performance on the MOSI dataset, verifying its flexibility to the more-than-two modalities case.

Comparison in large dataset. Considering existing datasets used in the experiments are relatively simple, we further conduct experiments on the UCF101 dataset which consists of 101 categories. The results are shown in the right part of Table 2. We can see that ARL achieves the best result, verifying its generalizability to more complex scenes.

4.4. Ablation Study

Generalization to different architectures. Apart from the convolutional neural network with the late-fusion method, transformer-based backbones using cross-modal interaction modules are also extensively applied. To validate the applicability of ARL in more scenarios, we combine it with two intermediate fusion methods MMTM [17] and mmFormer [41] for all the datasets. Both of them fuse information within the modality encoder. Specifically, MMTM is a CNN-based architecture that integrates intermediate feature maps from different modalities using squeeze and excitation operations. mmFormer is a transformer-based architecture, that fuses the intermediate feature maps of different modalities with the cross-attention strategy. For both methods, we use only one frame for each video on each dataset to align their input on the benchmark. Besides, we unified the backbone network as ResNet18.As illustrated in Table 3, the proposed ARL demonstrates significant performance gains across various architectures, even when fusion is employed during encoder processing, suggesting its applicability in complex scenarios.

Generalization to larger backbone networks. Considering that ResNet18 is a relatively small backbone network,

Dataset	CREMA-D	AVE	KS
Method	Acc	Acc	Acc
MMTM	51.86	51.52	56.70
MMTM†	**58.68**	**60.88**	**63.72**
mmFormer	59.69	60.41	64.72
mmFormer†	**68.33**	**67.14**	**68.51**

Table 3. Performance on different datasets with CNN-based and transformer-based architectures, respectively. † indicates ARL strategy is applied, which achieves better performance.

Dataset	CREMA-D	AVE	KS
Vanilla	77.66	84.52	82.31
ARL	**80.01**	**91.11**	**85.88**

Table 4. Experiments of Swin-Base backbone network on different datasets. The fusion method is Gated fusion. 'Vanilla' means conventional multimodal model. 'ARL' means the multimodal model using gradient decoupling learning.

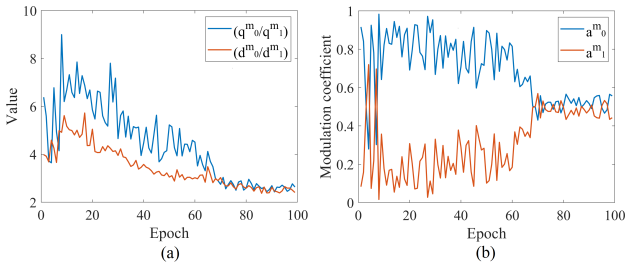

Figure 3. Visualization of the optimization dependency and modality variance ratio (a) and the corresponding modulation coefficient(b) of ARL on AVE during training. Here m_0 and m_1 represent audio and visual modalities, respectively.

Setting			Metric	
UR	AL	GR	Acc	macro F1
			58.83	59.43
✓			67.20	67.68
	✓		70.23	70.79
✓	✓		74.45	75.98
✓	✓	✓	**76.61**	**77.14**

Table 5. The ablation study of unimodal regularization (UR), asymmetric learning (AL), and gradient residual (GR) on the CREMA-D dataset. ✓ means the component is used.

we switch to Swin-Base to examine the scalability of ARL to a larger backbone. We initialize the model with the pre-trained weights on ImageNet-22k. As shown in Table 4, the introduction of ARL can continue to boost the performance of the vanilla model on all datasets. This demonstrates the effectiveness of ARL to large backbone networks.

The effect of each component in ARL. We conduct experiments on the CREMA-D dataset to study the effect. The results are shown in Table 5. We can see that all the unimodal bias regularization (UR), asymmetric learning (AL), and gradient residual (GR) significantly improve performance over the vanilla model. UR reduces decision bias, AL reduces decision variance, and GR ensures the gradient lower bound of the suppressed modality.

4.5. Visualization

To understand the mechanism of ARL intuitively, we visualize the difference between optimization dependency and modality variance ratio, and the corresponding modulation coefficient during the training process of the AVE dataset. As shown in Fig. 3 (a), the optimization dependency is lower than the modality variance ratio in the early training stage (before the 60th epoch). Therefore, as shown in Fig. 3 (b), ARL will give m_0 (the audio modality) a large weight to accelerate its optimization, increasing the optimization dependency on m_0 to increase the $\frac{d^{m_0}}{d^{m_1}}$. Then as shown in Fig. 3 (a), the curve of optimization dependency and modality variance will gradually converge to the same value. More importantly, we can see that the optimal ratio of optimization dependency is not 1 but 3, demonstrating the necessity of imbalanced learning.

5. Discussion

Conclusion. In this paper, we re-analyze the under-optimized problem in multimodal learning and reveal that balanced learning is not optimal for multimodal learning. In contrast, we prove that imbalanced learning obeying the inverse of modality variance can contribute to optimal performance. Then we propose a simple but effective multimodal learning strategy called Asymmetric Representation Learning (ARL) to alleviate the under-optimized problem, accelerating their optimization and encouraging models to rely on each modality according to the inverse of their variance. This method achieves consistent performance gain on five representative multimodal datasets under various settings. In addition, ARL can also generally serve as a flexible plug-in strategy for both CNN and Transformer-based models, demonstrating its practicality.

Limitation. To determine the variance of each modality, we need to measure the uncertainty of its prediction output. The current estimation of uncertainty via the self-information entropy is computationally efficient yet may be imperfect since it only leverages the information of a single inference. Intuitively, more accurate uncertainty estimation can lead to more accurate modulation coefficients. However, how to balance the complexity and accuracy of the uncertainty estimation is a challenging problem. We leave this problem to future work.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 1, 2

[2] Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20208–20217, 2022. 1, 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 6

[4] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022. 2

[5] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 6

[6] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021. 2

[7] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 2, 3

[8] Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3975–3984, 2021. 2

[9] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 1, 3

[10] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 1, 3, 6, 7

[11] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 103–118, 2018. 2

[12] Neha Gupta, Jamie Smith, Ben Adlam, and Zelda E Mariet. Ensembles of classifiers: a bias-variance perspective. *Transactions on Machine Learning Research*, 2022. 5

[13] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[14] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 2

[15] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*, 2024. 1, 3

[16] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3376–3390, 2021. 1, 2

[17] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020. 7

[18] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023. 1, 3, 6, 7

[19] Xiao Li, Lin Lei, Yuli Sun, and Gangyao Kuang. Dynamic-hierarchical attention distillation with synergetic instance selection for land cover classification using missing heterogeneity images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 2

[20] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015. 6

[21] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 1, 3, 6, 7

[22] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021. 2

[23] Mohsen Shahhosseini, Guiping Hu, and Hieu Pham. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 7:100251, 2022. 4

[24] Qiya Song, Bin Sun, and Shutao Li. Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1, 2

[25] 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. 6

[26] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 1407–1417, 2021. 1, 2

[27] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 6

[28] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 1, 2, 3, 6

[29] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. 3

[30] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. 2

[31] Shicai Wei, Chunbo Luo, Yang Luo, and Jialang Xu. Privileged modality learning via multimodal hallucination. *IEEE Transactions on Multimedia*, 2023. 1

[32] Shicai Wei, Yang Luo, Xiaoguang Ma, and Ren Peng. Mshnet: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 2

[33] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling. In *European Conference on Computer Vision*, pages 38–54. Springer, 2025. 2

[34] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024. 1, 3, 6

[35] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multi-modal cooperation via fine-grained modality valuation. *arXiv preprint arXiv:2309.06255*, 2023. 1, 3

[36] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86. Springer, 2025. 1, 3, 6

[37] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[38] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 6

[39] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024. 5

[40] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. *arXiv preprint arXiv:2311.10707*, 2023. 1, 3, 6

[41] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *arXiv preprint arXiv:2206.02425*, 2022. 7