CMoB: Modality Valuation via Causal Effect for Balanced Multimodal Learning

Jun Wang[1], Fuyuan Cao[1,2]*, Zhixin Xue[1], Xingwang Zhao[1], Jiye Liang[1]
[1]School of Computer and Information Technology, Key Laboratory of Computational
Intelligence and Chinese Information Processing of Ministry of Education,
Shanxi University, Taiyuan, China
[2]Shanxi Taihang Laboratory, Taiyuan, China
`jwang8532@gmail.com`, `cfy@sxu.edu.cn`, `xuezhixin@sxu.edu.cn`,
`zhaoxw@sxu.edu.cn`, `ljy@sxu.edu.cn`

Abstract

Existing early and late fusion frameworks in multimodal learning are confronted with the fundamental challenge of modality imbalance, wherein disparities in representational capacities induce inter-modal competition during training. Current research methodologies primarily rely on modality-level contribution assessments to measure gaps in representational capabilities and enhance poorly learned modalities, overlooking the dynamic variations of modality contributions across individual samples. To address this, we propose a **C**ausal-aware **Mo**dality valuation approach for **B**alanced multimodal learning (CMoB). We define a benefit function based on Shannon's theory of informational uncertainty to evaluate the changes in the importance of samples across different stages of multimodal training. Inspired by human cognitive science, we propose a causal-aware modality contribution quantification method from a causal perspective to capture fine-grained changes in modality contribution degrees within samples. In the iterative training of multimodal learning, we develop targeted modal enhancement strategies that dynamically select and optimize modalities based on real-time evaluation of their contribution variations across training samples. Our method enhances the discriminative ability of key modalities and the learning capacity of weak modalities while achieving fine-grained balance in multimodal learning. Extensive experiments on benchmark multimodal datasets and multimodal frameworks demonstrate the superiority of our CMoB approach for balanced multimodal learning.

1 Introduction

Humans construct multi-dimensional perception through multiple sensory modalities like vision, touch, hearing, and smell, processing information hierarchically to understand the real world [1, 2, 3]. This biological cognitive mechanism has inspired the development of multimodal learning paradigms in the field of machine learning. These paradigms involve constructing collaborative learning frameworks across heterogeneous modal representation spaces (e.g., text, images, audio, and video) to simulate the cognitive process of human cross-modal information fusion [4]. From a cognitive neuroscience perspective [5, 6, 7], this paradigm aligns with the neural mechanisms of the human brain's multisensory integration, where different sensory cortices enable collaborative learning across heterogeneous modalities through synaptic plasticity.

In recent years, although extensive research on multimodal learning has made significant progress, some studies have found that most existing multimodal models encounter the challenge of modality

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

collapse during joint training. This is because deep neural networks tend to prioritize modalities that are easy to learn while neglecting other modalities, thus failing to effectively integrate heterogeneous cross-modal information [8, 9, 10]. This issue prevents the model from fully leveraging the complementary information across all modalities. Consequently, this information loss results in the performance of unimodal in joint multimodal training falling far short of their intrinsic performance ceilings. Researchers define this phenomenon as multimodal imbalance learning [11, 12, 13, 14, 15]. The primary cause of this phenomenon stems from the greedy nature of deep neural networks: these networks tend to rely on high-quality modality that is highly relevant to the target task, thereby inhibiting the optimization of other modalities. In the joint training process of multimodal learning, this greediness makes the modality dominant in the joint training, inhibiting the learning of other modalities and generating a situation of modal competition. Researchers define this phenomenon as multimodal imbalance learning [16, 17].

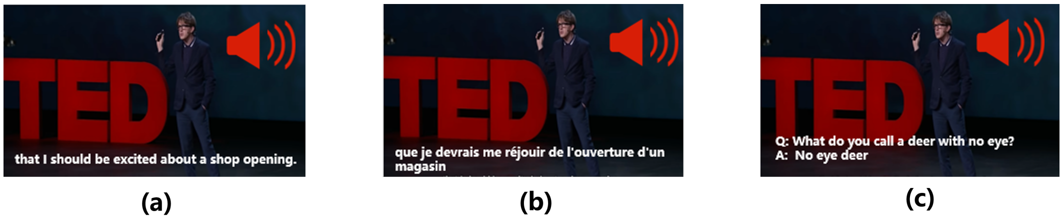

that I should be excited about a shop opening.
que je devrais me réjouir de l'ouverture d'un magasin
Q: What do you call a deer with no eye?
A: No eye deer
(a)
(b)
(c)

Figure 1: A TED talk with audio, video, and text modalities.

Recent research has focused on mitigating inter-modal discrepancies and enhance modality-specific data utilization during training to address the above challenge for improved multimodal model efficacy. Several methods have been proposed to identify and improve the training of poorly learned modalities by controlling gradient variations, modality contribution and incorporating loss functions [15, 18, 19, 20]. Gradient-based methods for assessing modality contribution typically rely on the assumption that larger gradients indicate higher modal importance. However, gradient values are susceptible to fluctuations from model parameters and training data, leading to unstable evaluations. Even though all modalities of a sample express the same concept (i.e., share identical label information), the amount of information related to the target object or target event varies across different modalities, which leads to differences in the data quality of individual modalities. These methods are only limited to modality-level analysis and fail to capture contribution differences at the sample level. In many real-world datasets, not all samples of image modal data, video modal data are more informative than text modal data. Take a TED talk video as an example (as shown in Figure 1): for video comprehension, the speech stream (audio information) and subtitle stream (text information) of a TED talk convey consistent core information. Viewers typically adopt intuitive, information-rich audio information as their primary source, parsing the speaker's speech stream to understand the meaning of the conveyed content. But when the target language suddenly exceeds the receiver's comprehension scope ((e.g., during a non-native language talk), the subtitle text replaces the auditory channel and becomes the core cognitive pathway for understanding the expression. Assessing the contribution degree of modalities at the modal level fails to reflect the true nature of the data. And although current encoders are effective in extracting multimodal features, the lack of interpretability [21, 22, 23] of the depth model makes it difficult to observe the role of each modality in the final prediction and to adjust unimodal training accordingly.

Inspired by the multimodal cognitive mechanisms of the human nervous system, we address the above problem by analyzing the learning process of the human brain's integrated processing of heterogeneous multimodal data [5, 7, 24, 25, 26]. We illustrate this concretely using a TED talk video example, as shown in Figure 1(d). When the speaker utters "no eye deer"—a homophonic phrase identical in pronunciation to "no idea"—viewers relying solely on the audio stream may struggle to grasp the intended meaning. In such cases, the cognitive system leverages instantaneous information entropy [27] to dynamically integrate subtitle data for assisted comprehension. During this process, the original cognitive system first extracts discriminative causal features [28] from audio data and based on experience to obtain the information entropy value of this modality. It then dynamically selects text modality data to assist in comprehension, thereby continuously refining the cognitive system's understanding of these modalities. We analyze this process and find that it has two core characteristics: (1) modality contribution valuation based on causality. By acquiring discriminative feature information and then evaluating modality contribution according to its information entropy, rather than simple signal quality comparison; (2) granular adjustment of the sample level. Even if a modality is overall reliable in the temporal dimension, the cognitive system will still implement cross-modality switching when an information bottleneck occurs in a single sample, and re-optimize its cognition of the modality. Inspired by this cognitive mechanism, in this paper we revisit the process of multimodal learning at the sample level from the perspective of causal learning. Causal learning aims to explore the essential causal relationship between things and reveal the real generation mechanism inside the data. We evaluate the degree of modality contribution by measuring the causal effect between samples, distinguishing the role of each modality in the final prediction and identifying low contributing modal samples. During the model training process, we enhance and optimize the low contributing modal samples to selectively improve the learning direction of the multimodal model to alleviate the modality imbalance problem and improve the performance and interpretability of multimodal learning.

We propose a causal-aware modality valuation method to evaluate the sample-level modality contribution in multimodal training. We design an optimization strategy for modality selection at the sample-level according to the contribution degree of each modality in order to mitigate the modality imbalance problem. In our experiments, we validate the effectiveness of our method by comparison and ablation experiments on publicly available data. In summary, our main contributions are as follows: (1) We evaluate the importance of sample in multimodal learning by means of a benefit function designed by information uncertainty theory. (2) We propose methods to quantify the degree of contribution of each modality from a causal perspective and represent the contribution of modalities in terms at the sample-level. (3) We propose a modality balancing approach from a data perspective to improve the performance of multimodal learning by optimizing the selection of weak modalities from the sample level in terms of modality contribution. (4) We validate the effectiveness and benefits of our approach by conducting extensive experiments on three publicly available datasets for different modalities compared to existing work.

2 Related work

2.1 Imbalance multimodal learning

The development of multimodal learning has been a great success, but many recent studies have found that the unimodal capabilities are not fully exploited in multimodal learning [1, 29, 30, 31]. Due to the imbalance in modality contributions during training, some unimodal capabilities are inhibited by the dominant modality and are unable to exert their upper limits [2, 8, 10, 17, 11, 15]. In some special cases, the performance of multimodal algorithms can even be lower than the performance of unimodal algorithms. Some researchers mitigate modality imbalance by analyzing the causes of modality imbalance and by modifying the process of forward propagation of model inference. For example, OPM algorithm [14]adopts a dynamic feature discarding approach for modality during multimodal training. OGM [17]achieves balanced multimodal learning by suppressing the dominant modality through an adaptive gradient adjustment strategy. PMR [12] accelerates the learning of weak modalities through category prototyping. GBlending [32]and MMpareto [16]reduce gradient conflicts during training by introducing unimodal assisted learning to control the direction of gradient update. Based on the above analysis, although these methods alleviate the modality imbalance to a certain extent, their approaches measure the degree of modality contribution through the modality-level and lack interpretability. Many scholars seek to investigate this issue from a sample-level perspective. SMLS [33] employs KL divergence to align the factual contribution distribution with the utopia contribution distribution. PDF [34]aims to reduce reliance on low-quality modalities via Relative Calibration (RC). The shapley value-based method [8], which enhances weak modalities by calculating combinations of all possible subsets. In this paper, we propose an interpretable method to quantify the degree of contribution of each modality in multimodal training. We characterize the degree of modality contribution in terms at the sample-level to mitigate the modality imbalance issue.

2.2 Causal learning

Causal learning aims to discover causal relationships from data and utilize these relationships to make interventions, predictions and decisions [35]. Causal learning focuses on identifying causal relationships, which is different from traditional machine learning that only discovers correlations between variables. Causality strictly distinguishes between "cause" and "effect" variables, and plays an irreplaceable role in revealing the mechanism of occurrence and guiding intervention behavior [36, 37, 38]. Causal inference has contributed to the development of artificial intelligence [26, 39, 40, 41] due to its ability to eliminate the harmful bias of confounders and to discover causal relationships among multiple variables. Causal effect are a central goal of causal inference, aiming to quantify the impact of interventions on outcomes by comparing the outcomes of treatment and control groups. Specifically, causal effect reveal causal relationships between variables by analyzing differences between potential outcomes across treatment conditions [28]. The central question is to answer: for a given individual or group, how do outcomes change when a certain intervention is applied compared to not applying that intervention? In this paper, we revisit the challenge of modality imbalance in multimodal joint training from a causal learning perspective.

3 Methodology

This section presents our method. First, we formulate the multimodal learning problem and its representation paradigm. Then, we propose an information-theoretic uncertainty measure for sample significance valuation and develop a causal-aware algorithm to quantify modalities contribution at the sample level. Finally, we propose a modality rebalancing approach from a data perspective to improve multimodal learning by optimizing the selection of weak modalities from the sample-level modality contribution.

3.1 Preliminary

With any loss of generality, We formalize the general formalism for expressing multimodal learning. Given a dataset $\mathcal{D} = \{(x_i^m, y_i^m) \mid i \in \{1, \cdots, N'\},\ m \in \{1, \cdots, M\}\}$is a finite and nonempty set for all modalities with N' data samples and M modalities. Each sample x has M modality, and y is the label of x can refer to a class, an answering, etc. $x_i = \{x_i^1, x_i^2, \cdots, x_i^M\}$, $y_i \in \{1, 2, \ldots, K\}$ denotes the corresponding class label from K classes. For ease of expression, we give examples with two modalities where v and a refer to the video and audio modalities, respectively. It is worth noting that our method is not limited to two modalities and can be extended to more modalities.

For multimodal joint training methods, we use the mainstream M modality-specific encoder to extract features from the original space. We use $h(\cdot)$ to define the feature extraction of the multimodal model. For any sample, the feature extraction process can be formalized as:

$$z_i^m = h(\theta^m, x_i^m),$$

where z denotes the d-dimension feature vector $z_i^m \in \mathbb{R}^d, \quad \theta^m$ denote the parameters of modality-specific encoder $\Theta^m(\theta^m, \cdot)$.

We obtain a joint representation of modality-specific features by fusing them through a fusion function $f(\cdot)$. Then, we use $\hat{F}(\cdot)$ to denotes the output function of the multimodal model, which maps vectors into $\mathbb{R}^K$, This procedure can be formally formulated as:

$$Z = f(z_i^1, z_i^2, \cdots, z_i^M),\ Z \in \mathbb{R}^D, \qquad \hat{F}(W, \theta, x, b) = W \cdot Z + b,$$

where D denotes the dimension of the joint feature representation, $W \in \mathbb{R}^{K \times D}$denotes the weight of the last layer of the forward propagation process, K denotes the number of categories in the task, and b is the bias term, $b \in \mathbb{R}^k$.

Finally, the cross-entropy loss is used for validation and the joint objective loss for multimodal learning can be formalized as:

$$L_{\text{cross-entropy}} = \frac{1}{N'} \sum_{i=1}^{N'} y_i \cdot \log(\operatorname{softmax}(\widehat{F})).$$

When updating the gradient for model backpropagation, W^m and the parameters of modality-specific encoder Θ^m are updated as:

$$W_{t+1}^m = W_t^m - \eta \frac{1}{N'} \sum_{i=1}^{N'} \frac{\partial L_{\text{cross-entropy}}}{\partial \hat{F}(x_i)} \cdot \Theta_i^m, \quad \theta_{t+1}^m = \theta_t^m - \eta \frac{1}{N'} \sum_{i=1}^{N'} \frac{\partial L_{\text{cross-entropy}}}{\partial \hat{F}(x_i)} \cdot \frac{\partial (W_t^m \cdot \Theta_i^m)}{\partial \theta_t^m},$$

$$\frac{\partial L_{\text{cross-entropy}}}{\partial \hat{F}(x_i)} = \frac{e^{\sum_{m=1}^{N'} W^m \cdot \Theta_i^m + b_{\hat{y}_i}}}{\sum_{k=1}^{K} e^{\sum_{m=1}^{N'} W^m \cdot \Theta_i^m + b_k}} - \mathbf{1}_{\hat{y}_i = y_i},$$

where η is the learning rate, $\hat{y}_i$ as the classification result of x_i. The phenomenon of modality imbalance can be formalized as: when one modality prediction has better outcome, its contribution $W^m \cdot \Theta_i^m$ dominates the logits output $\hat{F}(x_i, \cdot)$. This reduces the magnitude of $\frac{\partial L_{\text{cross-entropy}}}{\partial \hat{F}(x_i)}$, as the loss $L_{\text{cross-entropy}}$ already becomes smaller. The model is also optimized in the direction of smaller loss. Consequently, gradients for updating weaker modalities are suppressed, leading to under-optimized representations for them.

3.2 Sample benefit valuation

In multimodal learning, the multimodal model utilizes the different modalities of all samples for learning and inference, thereby enhancing the understanding of the data. For existing tasks of multimodal learning, all modalities are assumed to be predictive of the model [30]. The benefit function we designed is also based on this assumption. We mainly address the modality imbalance phenomenon of early fusion and late fusion in multimodal learning. We take the number of input modalities M as the value of the benefit function in multimodal joint learning when the prediction of the final network learning is matched with the ground truth.

In multimodal learning, the importance of samples changes as the model is trained iteratively. Due to the heterogeneity of modalities, the modality data itself also contains noise, inter-modal redundant information. In the initial stage of training, this may prevent the model from learning the high-level semantic relationship information of different modalities. In real scenarios, the data quality of different modalities of a sample is usually unstable, not all samples will have higher video modality than text modality [8]. In reality, multimodal models rely on different modalities for different samples, rather than depending on the same modality for all samples. Moreover, for different tasks in the same dataset, the amount of information related to different target tasks in the same sample is different, so it is difficult to accurately define the "quality" of modality in each sample.

According to Shannon's theory of information uncertainty [42], "the essence of information is to eliminate uncertainty" brings us new thinking. In other words, adding more modalities to a sample is accompanied by enhancing more information, which will bring less uncertainty to the multimodal model. We can approximate this relationship as follow:

Proposition 1. Given a sample has M modalities denoted as $x^{(M)} = \{x^1, x^2, \cdots, x^M\}$, assume that there any two subsets $x^{(B)}, x^{(C)}$ of $x^{(M)}$, and $x^{(B)} \subseteq x^{(C)} \subseteq x^{(M)}$, then for any multimodal classifier $\hat{F}(\cdot)$, it should be guaranteed that

$$\text{Conf}(\hat{F}(x^{(B)})) < \text{Conf}(\hat{F}(x^{(C)})) \leq \text{Conf}(\hat{F}(x^{(M)})).$$

Current encoders and classifiers have achieved significant advancements. For any reliable multimodal classifier, when new modalities are added, the prediction confidence of multimodal joint learning should increase and exhibit a positive correlation with the addition of modalities.

We defined a benefit function to evaluate the importance of the samples in the multimodal model learning process as follow:

$$B(M) = \begin{cases} |M|, & \text{if } \hat{F} = y_i, \ \text{Conf}(\hat{F}(x^{(C)})) \leq \text{Conf}((\hat{F}(x^{(M)}), \ \text{and } C \leq M, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

When the multimodal model's prediction is correct and the confidence after adding new modalities is higher than that before their addition, in this case the benefit of this sample is defined as the number of input modalities.

3.3 Causal-aware quantification method

In this subsection, we propose a causal-aware modality contribution quantification method from a causal perspective to capture fine-grained changes in modality contribution within samples. We dynamically evaluate the contribution degrees of different modalities in samples by calculating the changes in causal effects due to intervention modalities.

In the previous subsection, we evaluated the benefit value of the sample to the target task and measured the importance of the sample through the benefit function. Cognitive science research has revealed a unique mechanism in multimodal information processing in which the human modality choice behavior has a significant causal inference property [5]. This characteristic suggests that humans do not simply integrate all available information when faced with complex multimodal environments, but selectively focus on and utilize modality-specific information based on judgments of causality and causal strength.

Intervention is used as a straightforward and effective method to assess the existence of direct causal relationships between events [28, 35, 39]. Inspired by the concept of intervention in causality, we assess the effect of one event (cause) leading to the occurrence of another event (outcome) by means of intervention, and quantify the degree of contribution of modality in the sample. For multimodal deep neural networks, we observe the causal relationship between the variable t_i of the input set T and its corresponding output $H(T)$ through an intervention denoted as $do(t_i = x)$, as shown in Eq.2

$$\phi_H[t_i] = \mathcal{V}(H(T)) - \mathcal{V}(H(T|do(t_i = x))) = \mathcal{V}(H(T)) - \mathcal{V}(H(T')). \tag{2}$$

Here, $H(T)$ denotes control group, $H(T')$ denotes the treatment group. The difference between these two refers to the effect of the intervention. $\phi_H[t_i]$ can be regarded as the ITE(Individual Treatment Effect) of t_i through the function $H(\cdot)$. $\mathcal{V}(\cdot)$ can be any function that valuates the effect of an outcome. T' denotes the intervened outcome.

For each sample which contains M heterogeneous modalities, $S(x_i) = \{x_i^1, x_i^2, \cdots, x_i^j, \cdots x_i^M\}$ denotes the set of modalities of the samples. We want to measure the ITE of modality j in sample i. The modality of the treatment variable is denoted as x_i^j, the control group is $S(x_i)$, and the treatment group is $S(x_i)\backslash x_i^j$ then its individual causal effect can be expressed as follows:

$$\begin{aligned} ITE(x_i^j) = \phi_H[x_i^j] &= \mathcal{V}(H(T)) - \mathcal{V}(H(T')) \\ &= \mathcal{V}(H(S(x_i))) - \mathcal{V}(H(S(x_i)\backslash x_i^j)) \\ &= B(\hat{F}(S(x_i))) - B(\hat{F}(S(x_i)\backslash x_i^j)) \\ &= B(\hat{F}(S(x_i))) - B(\hat{F}(S(x_i)|do(t_i = x_i^j))), \end{aligned} \tag{3}$$

where $B(.)$ is the benefit function presented in the previous subsection. $\hat{F}(.)$ denotes is the output function of the multimodal model. When quantifying the contribution of the modality j in the sample i, we must not only calculate the effect of the intervention modality j on the output but also consider its own impact on the output. Thus, its contribution can be expressed as follows:

$$\begin{aligned} \Phi(x_i^j) &= ITE(x_i^j) + \mathcal{V}(H(x_i^j)) \\ &= B(\hat{F}(S(x_i))) - B(\hat{F}(S(x_i)|do(t_i = x_i^j))) + B(\hat{F}(x_i^j)). \end{aligned} \tag{4}$$

Additionally, as specified in Eq.1, when the multimodal model makes an incorrect prediction, the control group is $B(\hat{F}(S(x_i)))$=0. The minimum of treatment group $B(\hat{F}(S(x_i)|do(t_i = x_i^j)))$ is 0, and $B(\hat{F}(x_i^j)) \leq$1. Then we have $\Phi(x_i^j) \leq 1$, at which point we consider this treatment is not positive and the modality j in sample i is weak modality. We conducted further discussions on ITE, at Appendix A.6.

3.4 Dynamic modality optimization strategy

After quantifying the modality contribution at the sample level, we obtain the contribution of different modalities for each sample, and then distinguish the weaker modalities. We propose Algorithm 1, which implements a dynamic enhancement strategy to optimize weaker modalities during training iterations. To alleviate modality imbalance problem, we design a specific optimization function Eq.5 to selectively enhance the samples of weak modalities. The enhancement of weak modalities inevitably induces augmented modality-specific data within samples, thereby increasing the likelihood of model overfitting during training. We employ adaptive masking to mitigate model overfitting. We mask localized features of specific modalities, thereby compelling the model to leverage cross-modal contextual dependencies to achieve optimization objectives. This method strengthens inter-modal semantic coherence while preventing over-reliance on partial characteristics of individual modalities.

```
Algorithm 1 : CModB Algorithm
Input: Dataset $\mathcal{D} = \{(x_i^m, y_i')|i \in \{1, \dots, N'\}, m \in \{1, \dots, M\}\}$(train / val / test), device, method.
Output: Learned parameters $\theta$ of multimodal model.
INIT: Optimized dataset $D^{op}$, number of modalities $M$, initial parameters $\theta^0$, maximum iterations $E$, learning rate $\eta$, freeze_train epoch $F$.
for $e = 1$ to $E$ do
    if $e < F$ then
        Update model parameters $\theta$ with dataset $\mathcal{D}$;
    else
        Update $\mathcal{D}^{op} := \mathcal{D}$;
        for each sample $(x_i^m, y_i')$ in $D$ do
            Encoder feature extraction $h(\cdot)$;
            Feature fusion $f(\cdot)$;
            calculate the each unimodal ITE using Eq.3;
            calculate the each unimodal contribution $\Phi(x_i^j)$ using using Eq.4;
            Obtain optimize status using Eq.5;
            Update dataset $\mathcal{D}$ by the optimize status to obtain an optimized dataset $\mathcal{D}^{op}$;
        end for
        Update model parameters $\theta$ with dataset $D^{op}$.
    end if
end for
```

By applying masking, we can simulate real-world scenarios where data may exhibit incompleteness or partial information loss. We employ Time Masking for audio modality and Spatial Masking for video modality. This approach enhances training data diversity while mitigating the model's sensitivity to localized noise. So that the multimodal model learns to deal with incomplete data during training, and then has better adaptability and robustness when facing various complex situations in the real world.

$$\mathrm{Re}(x^i) = \begin{cases} f_{\mathrm{Re}}(1 - \Phi(x_i^j)) \cdot \mathrm{Mask}(\Phi(x_i^j)) & \Phi(x_i^j) \leq 1, \\ 0 & \Phi(x_i^j) > 1, e <= E/2, \\ \alpha \cdot \mathrm{Mask}(1 - \Phi(x_i^j)) & \Phi(x_i^j) > 1, e > E/2. \end{cases} \tag{5}$$

Here, $f_{\mathrm{Re}}(.)$ is a monotonically increasing function and $\mathrm{Mask}(\cdot)$ is a mask function, e is the number of iterations, E is the maximum iterations of the multimodal model, α is a hyperparameter. We mainly focus on optimizing the weaker modalities with low contribution. We maintain the original $\mathcal{D}$ inputs for the other modalities during the first half of multimodal training, while applying masking to these during the latter phase. We employ time masking for audio modality and spatial masking for video modality We achieve directional and targeted modality rebalancing by the above method.

4 Experiments

4.1 Experimental setup

Datasets: We select five public datasets,including CREMA-D[43], Kinetic Sounds[44], UCF-101[45], CMU-MOSEI[46], and NVGesture[47] datasets to validate our proposed method. The description of the complete dataset is provided in Appendix A.1.

Implementation Details: Details of the implementation are given in the Appendix A.2.

Baselines: We compare a range of baseline methods. These contain traditional MML approaches and fusion methods for balanced multimodal learning, namely, feature concatenation (Concat), prediction summation (Sum), prediction weighting (Weight) [48], MMCosine [49], AGM [50], OGM [17], GBlending [32], PMR [12], MMCooperation [8], Relearning [51], MLA [10], MMPareto [16].

4.2 Comparison with imbalanced MML baselines

We conducted comparative experiments on multiple datasets with various modalities and compared with SOTA methods. The main results for all datasets are presented in Table 1, where "CMoB" denotes our proposed approach. In Table 1, Unimodal-1 denotes training models using only the audio modality for datasets like CREMA-D, Kinetics Sounds, and CMU-MOSEI. For UCF-101 and NVGesture, this configuration corresponds to the RGB modality. Unimodal-2 infers training exclusively on the video modality for CREMA-D, Kinetics Sounds, and CMU-MOSEI. For UCF-101 and NVGesture, it corresponds to the optical flow modality. Unimodal-3 applies to CMU-MOSEI, where models are trained using the text modality, and to NVGesture, where the depth modality is utilized. It is worth noting that **Concat** refers to the baseline method commonly used in multimodal learning for mitigating modality imbalance problem, which employs concatenation fusion with a single multimodal cross-entropy loss function. Based on the results, the following key observations can be drawn: (1) Our proposed CMoB approach demonstrates strong generalization ability by showing excellent consistent performance across diverse heterogeneous multimodal datasets. Compared with PMR and OGM, which are rebalancing methods only applicable to two modalities, our method can handle datasets with more modalities and achieve the best results. (2) Each of the modality rebalancing frameworks significantly outperform the baseline Concat method relying on direct feature concatenation, with consistent performance improvements observed across evaluation metrics. This empirical evidence substantiates the objective existence of multimodal imbalance challenges, necessitating dynamic modality-specific parameter adjustment during multimodal joint training. (3) In the CMU-MOSEI dataset, there is a clear phenomenon that the best unimodal performance (Unimodal-3) surpasses its multimodal learning counterpart and outperforms many well-performing rebalancing methods on other datasets. In CREMA-D and CMU-MOSEI dataset, conventional fusion methods without modality balancing exhibit only minimal performance gains when compared to the performance of the important unimodal (the dominant modality in multimodal training such as Unimodal-1 in CREMA-D, Unimodal-3 in CMU-MOSEI). We will discuss and visualize this in the next subsections.

Table 1: Comparison with state-of-the-art (SOTA) multimodal learning algorithms. Bold and underlined results denote the best and second-best performances, respectively.

Method	CREMA-D		KineticsSounds		UCF-101		CMU-MOSEI		NVGesture	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Unimodal-1	61.17	60.63	55.06	54.96	78.60	77.49	71.09	41.7	78.22	78.33
Unimodal-2	49.56	47.81	45.31	43.76	59.90	58.19	71.03	41.68	78.63	78.65
Unimodal-3	-	-	-	-	-	-	80.58	74.57	81.54	81.83
Concat	65.5	65.07	65.63	65.28	81.8	81.21	78.99	69.40	81.33	81.47
Sum	63.44	63.12	64.97	64.72	80.21	79.42	79.10	71.15	82.99	83.05
Weight	66.53	66.41	65.33	64.89	82.65	82.19	79.94	72.31	82.42	82.57
MMCosine	67.19	67.34	67.49	67.09	82.97	82.47	80.38	73.67	81.52	81.55
AGM	71.59	72.11	66.62	65.88	81.7	80.89	79.86	71.89	82.78	82.82
OGM	67.76	68.02	67.04	66.95	82.07	81.3	-	-	-	-
GBlending	71.59	71.72	68.82	66.43	85.01	84.5	79.64	73.29	82.33	82.91
PMR	67.19	67.20	67.11	66.87	81.93	81.48	-	-	-	-
MMCooperation	75.85	76.68	68.01	68.03	85.25	84.69	79.84	72.99	82.85	83.02
Relearning	71.02	71.46	65.92	65.48	82.87	82.15	78.75	70.02	82.87	82.94
ARL	74.19	74.63	68.40	68.75	85.12	84.41	-	-	-	-
MLA	79.43	79.90	69.05	68.75	85.38	84.84	78.65	70.02	83.73	83.87
MMPareto	76.87	77.35	**74.55**	**74.21**	85.3	84.89	81.18	74.64	83.82	**84.24**
CMoB	**79.75** ±0.27%	**79.98** ±0.38%	72.03 ±0.22%	71.74 ±0.32%	**86.82** ±0.27%	**86.21** ±0.34%	**81.24** ±0.19%	**74.97** ±0.26%	**84.06** ±0.14%	84.18 ±0.21%

4.3 Comparison in scarcely informative modality case

In existing modal rebalancing methods, the mainstream view is to regard modalities with poor prediction performance as weak modalities and conduct additional training during the unimodal balancing. In practical applications, certain modalities may contain extremely limited label-relevant information and instead contain more noise. If the multimodal model is forced to learn information from these modalities under such circumstances, it will memorize more noise, thereby leading to negative outcomes. To simulate this scenario, we follow this paper [51] and choose to carry out experiments on the CREMA-D dataset. The video modality in the CREMA-D dataset consists of standardized emotional performances in a controlled laboratory environment, with its features primarily focusing on facial details of the human face. Compared to its audio modality, it inherently contains less label-relevant information. We modified its audio modal by adding extra white Gaussian noise. The experimental results are visualized as shown in Figure 2 and Appendix A.4.

From the visualization results, we can observe that when the available information in the dataset is very limited, the performance of existing methods all declines. As shown in Figure 2 the Concat method, the obvious cluster overlap in the video modality of the concat method indicates that it may be difficult to distinguish all categories in the high-dimensional space, and the categories in the audio modality are relatively dispersed. Through a direct comparison with the concat method, it can be visually observed that the modality rebalancing method improves the model's discriminability to a certain extent. This further validates the necessity of modality imbalance learning. In the second row of Figure 2 and Figure 5, the comparison of the video modality shows that under such extreme conditions, all methods have drawbacks. Their data points fail to form distinct separated clusters in the low-dimensional space. Our method shows relatively better performance in the video modality. In the audio modality, our method shows that the data points of the same category are more compact, while those of different categories are more separated. This indicates that our method makes it easier to distinguish these emotional categories. This proves that our method can also exhibit good robustness in such extreme scenarios.

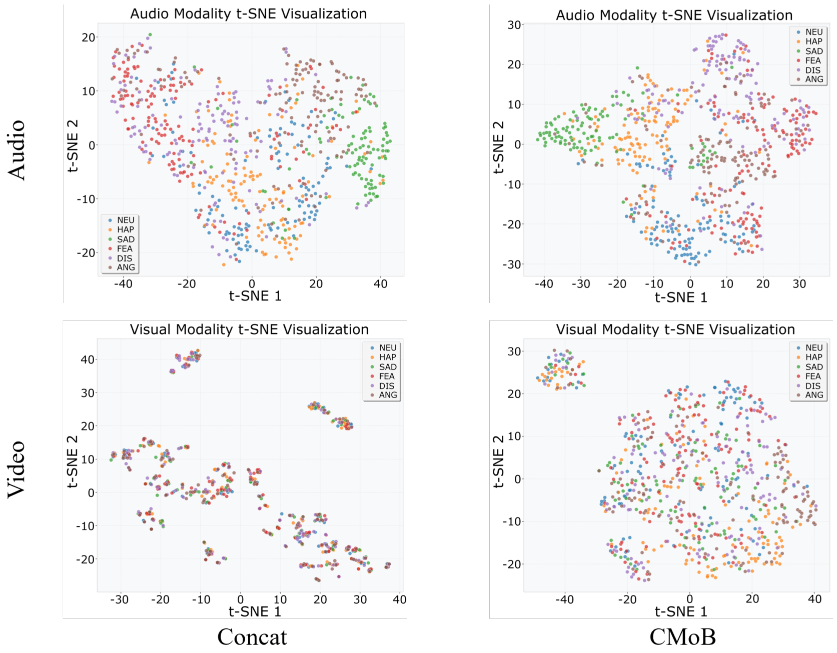


Figure 2: Each Unimodal representation visualization by t-SNE on the processed **CREMA-D** dataset. The six categories are indicated in different colors.

4.4 Ablation study

To comprehensively assess the effectiveness of our proposed method, we conduct experiments to study the influence of main components. Here, CQM represents our proposed a causal-aware modality contribution quantification method, and RE denotes the dynamic modality optimization strategy. We conduct an ablation study on CREMA-D and KineticsSounds datasets. The results are shown in Table 2. The implementation of the dynamic modality optimization method (RE) is predicated on the causal-aware modality contribution quantification method (CQM). Consequently, RE is inapplicable in the absence of CQM. From Table 2, we can see that both CQM and RE can boost performance in multimodal learning. Moreover, by integrating CQM with RE, the performance gap between audio modality and video modality is greatly reduced. Ablation studies clearly demonstrate the critical contribution of our proposed modules to overall method performance enhancement.

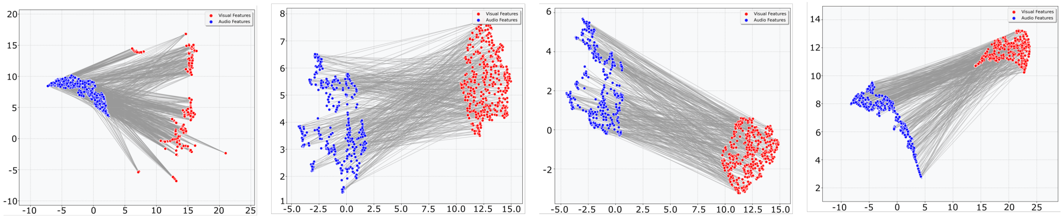
4.5 Further analysis

Analysis of Modality Gap: As mentioned in the paper [52], the existence of a geometric phenomenon in multimodal models: Modality Gap. It can denote regions in the shared space where the embeddings of different modalities are significantly separated, and a larger modality gap indicates better performance. We make further comparisons by validating the modality gap between our method and the better performing methods (which outperforms CMoB method with some data). The results are shown in Figure 3. In the concat method, the data points of the audio modality are over-clustered,

Table 2: Results of ablation study on CREMA-D and KineticsSounds datasets

Dataset	**Module**		**ACC**
	CQM	RE	
CREMA-D	×	×	65.50
	√	×	76.42
	√	√	79.75
KineticsSounds	×	×	65.63
	√	×	70.96
	√	√	72.03

while those of the video modality are over-dispersed, and a smaller modality gap. This indicates that the audio modality is primarily leveraged for discrimination in the Concat method. As shown in Figure 3(b) and Figure 3(c) , compared with MLA and MMPareto, our method has a larger modality gap, with more compactly clustered modality representations and fewer outliers, demonstrating its ability to learn more discriminative features and further validating its stability.



(a) Concat (b) MMPareto (c) MLA (d) CMoB

Figure 3: Visualizations of the modality gap on CREMA-D dataset.

Visualization: We employ Grad-CAM [53] to visualize the key regions focused by various rebalancing methods(Concat, MMPareto, MLA) during training on the CREMA-D dataset. Visualization results and analysis are provided in the Appendix A.3.

5 Conclusion

In this paper, we first analyze the limitations in existing modality rebalancing methods that neglect the dynamic variations of modality contributions at the sample level during training. Inspired by the remarkable causal inference power of the human modality choice behavior in cognitive science, we propose a causal-aware modality validation approach for balanced multimodal learning. We employ intervention methods to evaluate the causal effect, quantifying changes in modality contributions at the sample level during multimodal learning. The fine-grained evaluation approach enables targeted optimizations across modalities at the sample level, effectively mitigating the issue of multimodal imbalance. Comparative experimental results in multiple datasets validate the effectiveness of the proposed algorithm.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U24A20323 and 62376145), the Key Technologies Program of Taihang Laboratory in Shanxi Province (THYF-JSZX-24010700), the Science and Technology Innovation Talent Team of Shanxi Province (202204051002016), and the Taiyuan City "Double hundred Research action" of the first batch project about "Leading the Charge with Open Competition" (2024TYJB0127).

References

[1] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, “Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks,” in *Proceedings of the International Conference on Machine Learning*, pp. 24 043–24 055 , 2022.

[2] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, “Trusted multi-view classification with dynamic evidential fusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2551–2566, 2022.

[3] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal learning with transformers: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[4] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, “Multimodal biomedical ai,” *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.

[5] A. Occhipinti, S. Verma, C. Angione *et al.*, “Mechanism-aware and multimodal ai: Beyond model-agnostic interpretation,” *Trends in Cell Biology*, vol. 34, no. 2, pp. 85–89, 2024.

[6] H. Cui, A. Tejada-Lapuerta, M. Brbić, J. Saez-Rodriguez, S. Cristea, H. Goodarzi, M. Lotfollahi, F. J. Theis, and B. Wang, “Towards multimodal foundation models in molecular cell biology,” *Nature*, vol. 640, no. 8059, pp. 623–633, 2025.

[7] T. Zhong, J. Yu, Y. Pan, N. Zhang, Y. Qi, and Y. Huang, “Recent advances of platinum-based anticancer complexes in combinational multimodal therapy,” *Advanced Healthcare Materials*, vol. 12, no. 22, p. 2300253, 2023.

[8] Y. Wei, R. Feng, Z. Wang, and D. Hu, “Enhancing multimodal cooperation via sample-level modality valuation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27 338–27 347 , 2024.

[9] Y. Zhou, X. Liang, S. Zheng, H. Xuan, and T. Kumada, “Adaptive mask co-optimization for modal dependence in multimodal learning,” in *Processing of the International Conference on Acoustics, Speech and Signal*, pp. 1–5 , 2023.

[10] X. Zhang, J. Yoon, M. Bansal, and H. Yao, “Multimodal representation learning by alternating unimodal adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27 456–27 466 , 2024.

[11] Y. Yang, F. Wan, Q.-Y. Jiang, and Y. Xu, “Facilitating multimodal classification via dynamically learning modality gap,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 108–62 122, 2024.

[12] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, “Pmr: Prototypical modal rebalance for multimodal learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20 029–20 038 , 2023.

[13] Y. Zhou, X. Wang, H. Chen, X. Duan, and W. Zhu, “Intra-and inter-modal curriculum for multimodal learning,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 3724–3735 , 2023.

[14] Y. Wei, D. Hu, H. Du, and J.-R. Wen, “On-the-fly modulation for balanced multimodal learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] J. Fu, J. Gao, B.-K. Bao, and C. Xu, “Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4843–4856, 2023.

[16] Y. Wei and D. Hu, “Mmpareto: boosting multimodal learning with innocent unimodal assistance,” in *Proceedings of the International Conference on Machine Learning* , 2024.

[17] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, “Balanced multimodal learning via on-the-fly gradient modulation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247 , 2022.

[18] Y. Sun, S. Mai, and H. Hu, “Learning to balance the learning rates between various modalities via adaptive tracking factor,” *IEEE Signal Processing Letters*, vol. 28, pp. 1650–1654, 2021.

[19] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, “Audiovisual slowfast networks for video recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* , 2020.

[20] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, “Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably),” in *Proceedings of the International conference on machine learning*, pp. 9226–9259 , 2022.

[21] Z. You, Y.-H. Tsai, W.-C. Chiu, and G. Li, “Towards interpretable deep networks for monocular depth estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12 879–12 888 , 2021.

[22] P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux, “Model interpretability through the lens of computational complexity,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 487–15 498, 2020.

[23] F.-L. Fan, J. Xiong, M. Li, and G. Wang, “On interpretability of artificial neural networks: A survey,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 6, pp. 741–760, 2021.

[24] A. Khilkevich, M. Lohse, R. Low, I. Orsolic, T. Bozic, P. Windmill, and T. D. Mrsic-Flogel, “Brain-wide dynamics linking sensation to action during decision-making,” *Nature*, vol. 634, no. 8035, pp. 890–900, 2024.

[25] R. Rideaux, K. R. Storrs, G. Maiello, and A. E. Welchman, “How multisensory neurons solve causal inference,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 32, p. e2106235118, 2021.

[26] D. Mahajan, S. Tople, and A. Sharma, “Domain generalization using causal matching,” in *Proceedings of the International Conference on Machine Learning*, pp. 7313–7324 , 2021.

[27] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. Wiley, 2016.

[29] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[30] H. Ma, Q. Zhang, C. Zhang, B. Wu, H. Fu, J. T. Zhou, and Q. Hu, “Calibrating multimodal learning,” in *Proceedings of the International Conference on Machine Learning*, pp. 23 429–23 450 , 2023.

[31] J. Hu, J. Gu, S. Yu, F. Yu, Z. Li, Z. You, C. Lu, and C. Dong, “Interpreting low-level vision models with causal effect maps,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[32] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12 695–12 705 , 2020.

[33] Y. Zhou, X. Liang, Y. Xu, and X. Lin, “Dataset-aware utopia modality contribution for imbalanced multimodal learning,” *Information Fusion*, vol. 124, p. 103383, 2025.

[34] B. Cao, Y. Xia, Y. Ding, C. Zhang, and Q. Hu, “Predictive dynamic fusion,” in *Proceedings of the International Conference on Machine Learning* , 2024.

[35] K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio, “Interventional causal representation learning,” in *Proceedings of the International Conference on Machine Learning*, pp. 372–407 , 2023.

[36] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, “Causalvae: Disentangled representation learning via neural structural causal models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602 , 2021.

[37] S. Deshpande, K. Wang, D. Sreenivas, Z. Li, and V. Kuleshov, “Deep multi-modal structural equations for causal effect estimation with unstructured proxies,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 931–10 944, 2022.

[38] J. Zhang, E. Eldele, F. Cao, Y. Wang, X. Li, and J. Liang, “Counterfactual contrastive learning with normalizing flows for robust treatment effect estimation,” in *Proceedings of the International Conference on Machine Learning* , 2025.

[39] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, “Causality inspired representation learning for domain generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8046–8056 , 2022.

[40] J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen, “Weakly supervised causal representation learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 319–38 331, 2022.

[41] F. Cao, X. Jing, K. Yu, and J. Liang, “Fwcec: An enhanced feature weighting method via causal effect for clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 2, pp. 685–697, 2025.

[42] J. Soni and R. Goodman, *A mind at play: How Claude Shannon invented the information age*. Simon and Schuster, 2017.

[43] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[44] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617 , 2017.

[45] 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.

[46] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246 , 2018.

[47] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, “Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks,” in *Proceedings of the International Conference on Machine Learning*, pp. 24 043–2405 , 2022.

[48] Y. Yang, K.-T. Wang, D.-C. Zhan, H. Xiong, and Y. Jiang, “Comprehensive semi-supervised multi-modal learning.” in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4092–4098 , 2019.

[49] R. Xu, R. Feng, S.-X. Zhang, and D. Hu, “Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5 , 2023.

[50] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, “Boosting multi-modal model performance with adaptive gradient modulation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22 214–22 224 , 2023.

[51] Y. Wei, S. Li, R. Feng, and D. Hu, “Diagnosing and re-learning for balanced multimodal learning,” in *Proceedings of the European Conference on Computer Vision*, pp. 71–86 , 2024.

[52] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626 , 2017.

[54] Z. Lu, “A theory of multimodal learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 57 244–57 255, 2023.

[55] P. Grecov, K. Bandara, C. Bergmeir, K. Ackermann, S. Campbell, D. Scott, and D. Lubman, “Causal inference using global forecasting models for counterfactual prediction,” in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 282–294 , 2021.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose a causal-aware modality validation approach for balanced multimodal learning that captures the fine-grained changes in modality contribution degrees within samples.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitation of proposed method, CMoB, at Sec. 4.3

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions and proofs are clearly stated in Sec. 3 of our manuscript

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental settings are detailed in Sec.4.1 and Appendix A.2. All experiments can be easily reproduced with our code.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We make our code publicly available at `https://github.com/perpetual1859/CMoB` , including data and detail documentation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings and hyperparameter selection details are in Appendix A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See the results with standard deviation in Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the experiments compute resources at AppendixA.2. All experiments were run on 2 NVIDIA DGX A100.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, this paper conform the the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is a foundational research and has no direct negative social impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: No such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data and methods are explicitly mentioned.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methodology and experiments presented in this paper do not involve the use of large language models (LLMs) as an important, original, or non-standard component. No LLMs were employed for data processing, model design, training. The research was conducted independently of any generative or foundation model technologies; hence, an LLM declaration is not applicable.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Supplementary Materials and Experimental Details

A.1 Dataset

CREMA-D is an emotion recognition dataset with two modalities, audio and video. The dataset includes 7,442 clips annotated in six emotions. Kinetic Sounds is an action recognition dataset with two modalities, audio and video. This dataset contains 19k 10-second video clips categorized into 31 human action classes, which are selected from the Kinetics dataset. UCF-101 is an action recognition dataset with two modalities, RGB and optical fow. This dataset contains 10l categories of human actions with 9,537 samples in the training set and 3,783 samples in the test set. CMU-MOSEI is a sentiment analysis dataset with three modalities, audio, video, and text. This dataset includes more than 1,000 online speakers segmented into 23,453 utterances and is annotated with utterance-level sentiment labels. NVGesture is a gesture recognition dataset designed for human-computer interaction research and consists of three modalities, RGB, Depth, and optical flow. This dataset with 1,050 samples in the training set and 482 samples in the test set.

A.2 The implementation details

In our experiments, we use the raw data for experiments. Following [17, 12, 51], the architecture and initialization setup followed an unbalanced multimodal learning study for a fair comparison. For the CREMA-D and the Kinetic Sounds dataset, ResNet-18 is employed as the backbone for processing both audio and video data and trained from scratch. For the CMU-MOSEI dataset, we employ transformer-based networks as the backbone architecture, training the model from scratch. Encoders used for UCF-101 are ImageNet pre-trained. In term of video and optical flow modalities, we first select 10 frames from each clip and then uniformly sample three frames as input. We adjusted the input channels of ResNet18 from three to one to fit our data format. For audio modal data, we convert to a 257×299 spectrograms for CREMA-D and a 257×1004 spectrograms for KineticsSounds. For text-image datasets, our framework employs ResNet-50 as the image encoder and BERT for text processing, where images are resized to 224×224 resolution and text sequences are truncated to a maximum length of 128 characters. During training, we use the SGD optimizer with momentum (0.9) and set the learning rate at 1×10^{-3} . All models are trained on 2 NVIDIA DGX A100.

A.3 Visualization

We employ Grad-CAM [53] to visualize the key regions focused on by various rebalancing methods(Concat, MMPareto, MLA) during inference on the CREMA-D dataset. By computing gradients of target class scores with respect to the last convolutional feature map and generating pixel-level importance weights, Grad-CAM can precisely localize visual regions that significantly contribute to the decision-making process in the target task (emotion recognition). This visualization method not only helped us to verify the effective use of weak modality features by different models, but also provided an intuitive basis for us to analyze the allocation of visual attention in multimodal interactions. The visualization results are presented in Figure 4, where the first, second, third, fourth and last columns denote the results of the tenth, thirtieth, fiftieth and eightieth last epoch, respectively. This image is a keyframe extracted from the video modality of the category "NEU" in the CREMA-D dataset. Based on the heat map of the video modality, the following key observations can be drawn: (1) In the emotion recognition task, audio modality has richer information than the video modality, especially in the CREMA-D dataset. So we found that the video modality provides limited information at the early stage of training and relies mainly on the audio modality to dominate the feature extraction. At the 50th epoch of training, the video modality of the Concat method starts to focus on the feature extraction of facial expressions. However, the method is optimized by a uniform joint loss function, at this point, the dominant modality (audio modality) leads to a modality imbalance problem that causes the subsequent optimization of the video model to be directionally biased. (2) At different stages of model training, compared to other modal rebalancing methods our method focuses on the features of a person's face that are most relevant to the emotion recognition task, verifying the effectiveness and stability of our method in weak modality learning.

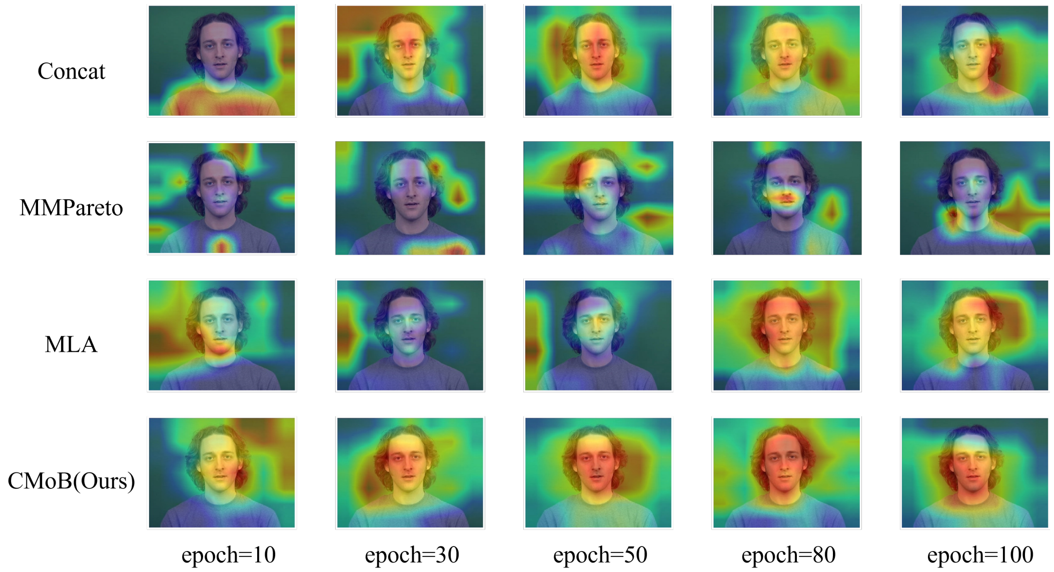


Figure 4: Visualizations of various rebalancing methods on CREMA-D dataset.

A.4 Comparison in scarcely informative modality case

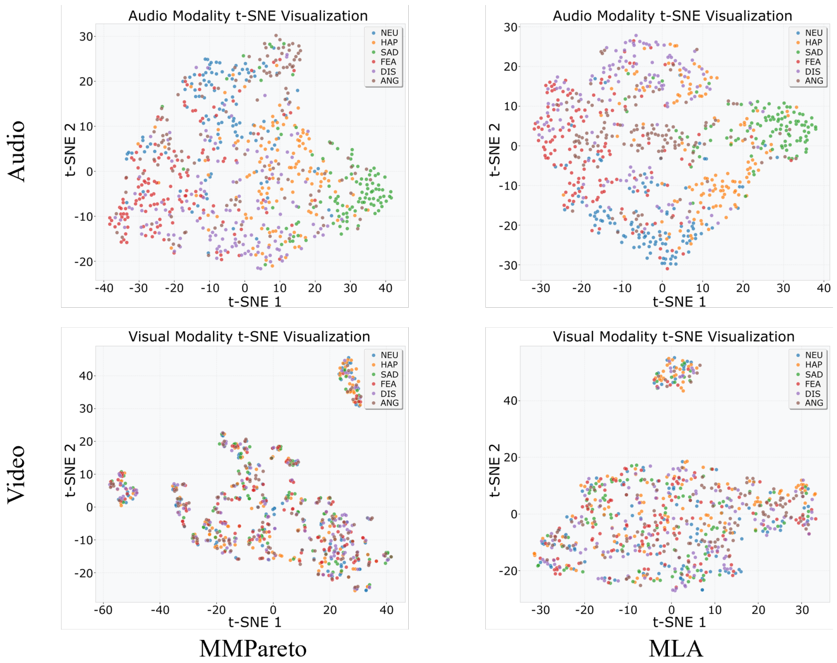


Figure 5: Each Unimodal representation visualization by t-SNE on the processed **CREMA-D** dataset. The six categories are indicated in different colors.

A.5 Further discussion on ITE

A further question we need to consider is: in multimodal learning, does our evaluation of the individual treatment effect (ITE) possess globality?

"Globality" generally refers to the property that an effect or conclusion holds universally among different individuals, groups, or environments, emphasizing its generalizability and cross-scenario applicability [54, 55]. Take classification tasks as an example: our goal is to train a robust multimodal deep neural network that constructs a generalizable decision boundary, accurately mapping input data into predefined classes. During training, all multimodal data also operate within the unified deep neural network. After multiple iterations, the resulting discriminative model can classify heterogeneous data from diverse modalities, and it has applicability in different modality in various samples. Therefore, we calculate that the individual causal effect of input modalities and output results processed by the multimodal deep network possesses globality.

A.6 Comparison with MML baselines on large-scale datasets

We conduct an experimental on the relatively large-scale dataset VGGSound. The VGGSound datasets consist of both audio and video modalities. The VGGSound dataset, which contains 310 classes and a wide range of audio events in everyday life, is a relatively large dataset. It includes 168,618 videos for training and validation, and 13,954 videos for testing. The experimental results show the superiority of our method. The results of the comparative experiment are shown in Table 3

Table 3: Comparison with different methods on VGGSound dataset.

Method	MAP	ACC
AGM	51.98%	47.11%
MLA	54.73%	51.65%
ReconBoost	53.87%	50.97%
MMPareto	54.74%	51.25%
Ours	**54.98%**	**51.74%**

To further validate the effectiveness of our method under scenarios with more modalities, we follow this paper and conduct further experiments on the Caltech101-20 dataset [8]. We compare our method with the Concat method and the Shapley value method. The experimental results confirm the effectiveness of our method, even when using five modalities (views), as shown in Table 4.

Table 4: Accuracy of our methods on Caltech101-20 dataset.

Num of modalities	Concat	Shapley	Ours
2	82.91	83.47	**83.71**
3	87.71	87.99	**88.22**
4	93.64	94.07	**94.35**
5	94.63	94.73	**94.86**