Improving Multimodal Learning Balance and Sufficiency through Data Remixing

Xiaoyu Ma [1 2] **Hao Chen** [1 2] **Yongjian Deng** [3]

Abstract

Different modalities hold considerable gaps in optimization trajectories, including speeds and paths, which lead to *modality laziness* and *modality clash* when jointly training multimodal models, resulting in insufficient and imbalanced multimodal learning. Existing methods focus on enforcing the weak modality by adding modality-specific optimization objectives, aligning their optimization speeds, or decomposing multimodal learning to enhance unimodal learning. These methods fail to achieve both unimodal sufficiency and multimodal balance. In this paper, we, for the first time, address both concerns by proposing multimodal Data Remixing, including decoupling multimodal data and filtering hard samples for each modality to mitigate modality imbalance; and then batch-level reassembling to align the gradient directions and avoid cross-modal interference, thus enhancing unimodal learning sufficiency. Experimental results demonstrate that our method can be seamlessly integrated with existing approaches, improving accuracy by approximately **6.50%**↑ on CREMAD and **3.41%**↑ on Kinetic-Sounds, without training set expansion or additional computational overhead during inference. The source code is available at Data Remixing.

1. Introduction

Multimodal learning (Ngiam et al., 2011) is a rapidly evolving field in artificial intelligence, aimed at enhancing the perception and decision-making capabilities of models by integrating data from diverse modalities, including vision, sound, and text (Zhu et al., 2024). However, existing multimodal learning methods often face challenges in fully integrating rich multimodal knowledge across different modalities. Due to the inherent differences in data representation and distribution across modalities, their optimization trajectories differ significantly, resulting in imbalanced learning when multiple modalities are jointly trained under a unified objective. Specifically, multimodal models often prioritize learning the most discriminative features from the strong modality, suppressing the training of the weak modality and causing it to become lazy, known as *modality laziness* (Wang et al., 2020; Huang et al., 2022; Du et al., 2021).

Several methods have been proposed to address these issues. Some focus on alleviating modality laziness by enforcing the weak modality through modality-specific optimization objectives, such as adjusting task-specific supervision (Wang et al., 2020; Xu et al., 2023; Du et al., 2023), introducing prototype learning (Fan et al., 2023), or using

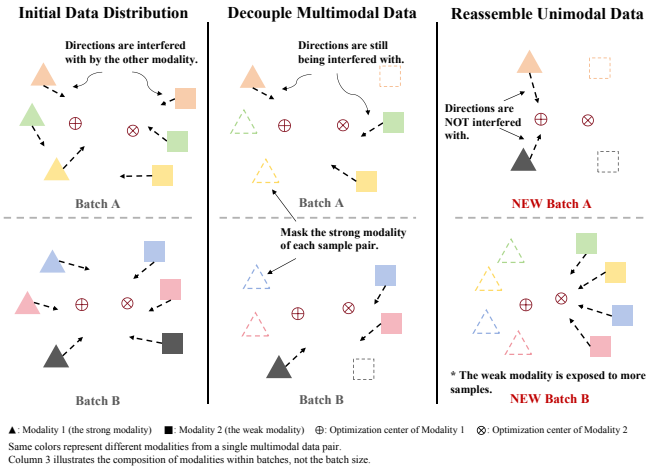


Figure 1. We decouple the multimodal data to assign samples to each modality's training and then reassemble the inputs to control the consistency of modalities within the batch. By regulating the number of samples, we mitigate modality laziness, and by adjusting the batch composition, we alleviate modality clash.

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China [2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China [3]College of Computer Science, Beijing University of Technology, Beijing, China. Correspondence to: Hao Chen <haochen303@seu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

knowledge distillation (Du et al., 2021). Other methods aim to align optimization speeds by modifying learning rates (Sun et al., 2021) and gradients (Peng et al., 2022; Li et al., 2023; Sun et al., 2023; Fu et al., 2023; Kontras et al., 2024; Wei & Hu, 2024) based on unimodal performance. In contrast, some approaches attempt to regulate modality inputs to more explicitly enhance the learning of weak modalities, such as by augmenting weak modality samples (Wu et al., 2022; Wei et al., 2024) and masking strong ones (Zhou et al., 2023), or by decoupling multimodal learning into unimodal tasks to avoid modality laziness (Zhang et al., 2024b). However, most of the existing methods align the optimization speeds of different modalities to alleviate modality laziness, without addressing the fact that even when modality balance is achieved, differing optimization paths can still cause interference during modality optimization. In multimodal learning, the gradient update directions of different modalities are inconsistent (Fan et al., 2023), leading to cross-modal interference during batch gradient descent (Figure 3(c)), which can be called *modality clash*. This inconsistency causes the modalities to deviate from their expected optimization paths, as shown in Figure 1, resulting in insufficient learning across all modalities and limiting the effectiveness of multimodal learning. Moreover, some methods are constrained by specific model architectures (He et al., 2022; Lin et al., 2023; Zhang et al., 2024b) or hinder training efficiency (Wei et al., 2024; Zhang et al., 2024b), which further limits their applicability.

Recognizing these limitations, we propose the **Data Remixing** method, illustrated in Figure 2. Through dynamic sample allocation and batch-level alignment mechanisms, we simultaneously address modality laziness and modality clash without hindering training efficiency or being constrained by specific model architectures.

Concretely, our Data Remixing method consists of two key steps: sample-level decoupling of multimodal data and batch-level reassembling of unimodal data. First, based on unimodal separability, we evaluate the representational capability of each modality at the sample level, retaining only the input from the weak modality while masking the others to zero. By decoupling the multimodal data, multimodal models can leverage specific samples to train each modality effectively. However, even after decoupling, unimodal gradient update directions still lead to interference. We further analyze the optimization process and propose that modality clash originates at the batch level. To address this, we reassemble unimodal data based on the decoupling assignments, ensuring that each batch contains data from only one modality, free from interference from other modalities, as shown in column 3 of Figure 1. Such sample-level decoupling and batch-level reassembling allow each modality to be learned sufficiently and balanced, all without expanding the dataset.

We test our method on different datasets and achieve excellent results. When combined with conventional fusion methods, the model's performance shows a notable improvement across multiple datasets. Importantly, our strategy does not expand the dataset or introduce additional overhead during inference. We summarize our key contributions as follows:

- We introduce the Data Remixing method, a training strategy that combines decoupling multimodal data and reassembling unimodal data to address modality laziness and clash. Our method does not require dataset expansion and is not constrained by specific model architectures.
- We analyze the optimization process and propose that modality laziness and clash originate at the batch level. To the best of our knowledge, we are the first to analyze the challenges of multimodal learning at the batch level and propose a solution.
- We perform experiments and demonstrate that (1) Data Remixing achieves excellent results in multimodal learning tasks; (2) Data Remixing can be easily and effectively integrated with other methods, leading to significant improvements of approximately **6.50%**↑ on CREMAD and **3.41%**↑ on Kinetic-Sounds.

2. Related Work

In this section, we introduce several strategies for multimodal balance learning. We categorize these methods based on the form of data input. The methods described in Section 2.1 retain multimodal inputs throughout the learning process, whereas those in Section 2.2 operate with unimodal inputs (though not necessarily for the entire duration of training).

2.1. Balance with Multimodal Joint Input

Wang et al. (2020); Du et al. (2021) discover that the optimization speed varies across different modalities. Therefore, when optimizing a multimodal model with a unified objective, the weak modality fails to learn adequately, leading to modality laziness. Most methods maintain multimodal data inputs during training, achieving modality balance by adjusting optimization objectives and speeds.

Some try to alleviate modality laziness by enforcing the weak modality through modality-specific optimization objectives Wang et al. (2020); Du et al. (2021); Xu et al. (2023); Du et al. (2023); Fan et al. (2023). Wang et al. (2020) propose Gradient-Blending, which calculates the optimal mixing mode of modality losses by determining the overfitting situation for each modality. Du et al. (2021) attempt to distill knowledge from well-trained unimodal models to enhance

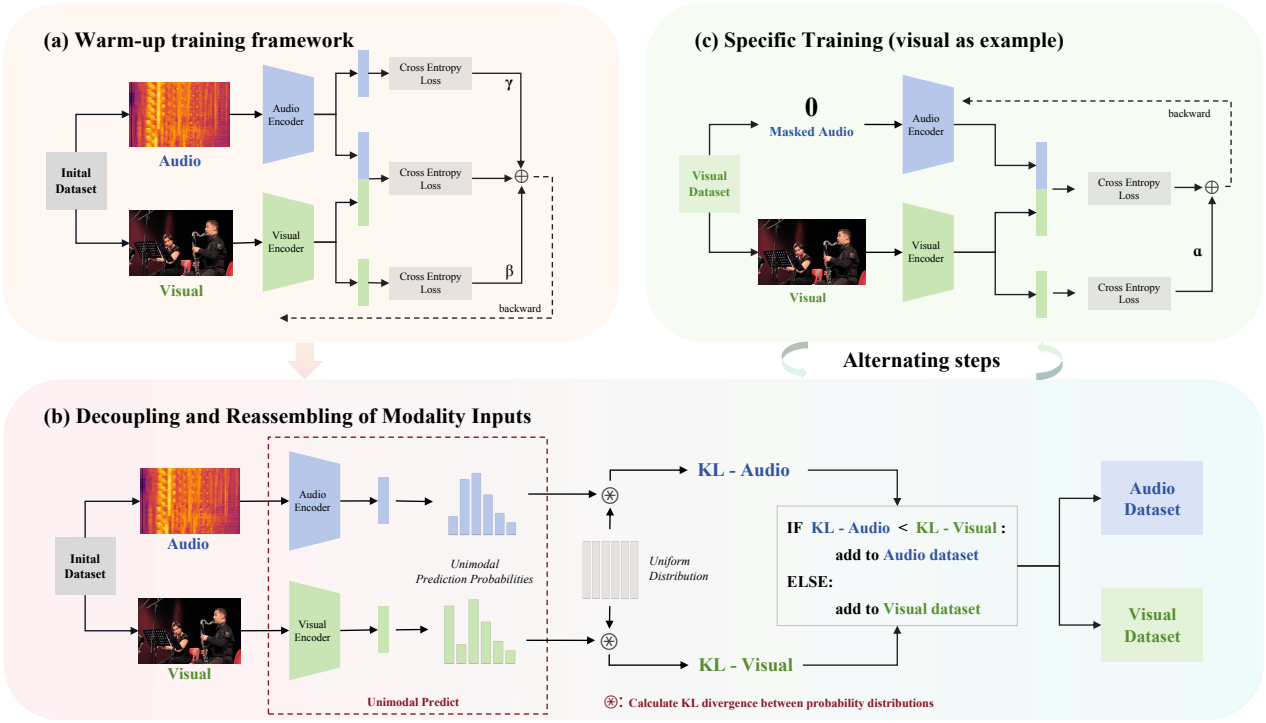


Figure 2. The pipeline of Data Remixing method. In step (a), the complete dataset is used for training to ensure the model develops the basic representational capability. In step (b), the multimodal data is decoupled based on unimodal separability (calculated using KL-divergence), and the original dataset is reassembled into non-overlapping subsets. In step (c), the subsets obtained in step (b) are used to train on the specific modality by masking other modalities to zero.

the unimodal encoders. Fan et al. (2023) introduce the prototype cross-entropy loss for each modality to accelerate the slow-learning modality. Others try to align optimization speed by changing learning rates (Sun et al., 2021) or modifying gradients (Peng et al., 2022; Li et al., 2023; Sun et al., 2023; Fu et al., 2023; Kontras et al., 2024) according to unimodal performance. Sun et al. (2021) dynamically adjust the learning rates of different modalities based on the unimodal predictive loss. Peng et al. (2022) adaptively modulate the gradients of each modality by monitoring the discrepancy in their contribution to the learning objective. These methods promote modality balance and enhance the expressive power of multimodal models. However, they fail to recognize that even when modality laziness is addressed, modality clash persists, meaning that multimodal capability remains limited.

2.2. Balance with Unimodal Single Input

Some methods attempt to use alternating or selective unimodal inputs to promote multimodal learning, thereby avoiding interference from other modalities and enhancing the expressive power of multimodal models (Wu et al., 2022; Zhang et al., 2024b; Wei et al., 2024; Zhou et al., 2023). Wu et al. (2022) measure the relative speed of each modality and train unimodal branches to fully utilize them. Zhang et al. (2024b) decompose multimodal learning into alternating unimodal learning with gradient modification to preserve cross-modal interactions. Wei et al. (2024) design a sample-level modality contribution evaluation based on Shapley values and perform resampling of the weak modality to enhance specific modality training. These methods, which rely on unimodal inputs, seem to avoid modality clash. However, upon closer inspection, we find that modality clash is introduced at the batch level. Like the methods in Section 2.1, they have not analyzed or addressed the root cause of modality clash.

Overall, current work focuses on addressing modality laziness but fails to recognize that modality clash still exists in balanced multimodal learning. Meanwhile, some methods are constrained by specific model architectures (Wu et al., 2022; Zhang et al., 2024b) or hinder training efficiency (Zhang et al., 2024b; Wei et al., 2024), limiting their applicability. In this paper, we aim to simultaneously mitigate modality laziness and modality clash while ensuring compatibility with various fusion methods and model architectures.

3. Method

3.1. Model formulation

In this paper, we focus on the multimodal discrimination task, following related works (Peng et al., 2022; Fan et al., 2023; Wei et al., 2024; Zhang et al., 2024b). For convenience, we consider two input modalities: m_a and m_v. The training dataset is $\mathcal{D} = \{x_i, y_i\}_{i=1,2,\ldots,N}$. Each x_i consists of multimodal inputs, i.e., $x_i = (x_i^a, x_i^v)$. $y_i \in \{1, 2, \ldots, M\}$, where M is the number of classes. The multimodal model is trained using batch gradient descent, with data $\boldsymbol{x} = (\boldsymbol{x}^a, \boldsymbol{x}^v)$ sampled from a batch B.

We use a multimodal model consisting of two unimodal branches for prediction. Each branch has a unimodal encoder, denoted as ϕ^a and ϕ^v, used to extract features from the corresponding modality of $\boldsymbol{x}$. The encoder outputs are represented as $\boldsymbol{z}^a = \phi^a(\theta^a, \boldsymbol{x}^a)$ and $\boldsymbol{z}^v = \phi^v(\theta^v, \boldsymbol{x}^v)$, where θ^a and θ^v are the parameters of the encoders. The results of the two unimodal encoders are fused in some way (Owens & Efros, 2018; Gunes & Piccardi, 2005) to obtain the multimodal output. We use cross-entropy (CE) loss as the loss function. To achieve better representation ability in the warm-up stage, we add a separate classification head for each modality, as Wang et al. (2020) do, and modify the loss function as follows:

$$\mathcal{L} = \mathcal{L}_{CE}^0 + \sum_{k=1}^{K} w_k \mathcal{L}_{CE}^{m_k}, \tag{1}$$

where $\mathcal{L}_{CE}^0$ represents the CE loss for multimodal prediction, while $\mathcal{L}_{CE}^{m_k}$ represents the CE loss for the individual prediction of the k-th modality. We provide experiments to demonstrate that the improvements brought by our method are not dependent on the design of this loss function.

3.2. Data Remixing Method

Overview of Method. The complete pipeline of our method is presented in Figure 2. In step (a), we first use the complete dataset and multimodal inputs for warm-up training to ensure the model has the basic representational capability. Then, the model is optimized through alternating steps (b) and (c). In step (b), we decouple the multimodal inputs based on the KL divergence of unimodal prediction probabilities and reassemble the data at the batch level according to the remaining modality. In step (c), we perform specific training for each modality using the reassembled dataset.

Decouple Multimodal Data. In multimodal learning, the model tends to fit the modality that optimizes faster, often converging before the other modality has been sufficiently learned (Du et al., 2021). Therefore, appropriately reducing the training speed of the strong modality may be an effective strategy for balancing learning (Peng et al., 2022). However, modality-level control over the optimization process or objectives overlooks the imbalance at the sample level between modalities (Wei et al., 2024), which limits the effectiveness of the balancing. Based on this premise, we consider promoting balance by assigning specific samples (in terms of difficulty and quantity) to different modalities, achieving a more precise and convenient modality balance.

We first define a sample-level unimodal capability evaluation method to achieve either-or data allocation, instead of using Shapley values as proposed by Wei et al. (2024)[1], to avoid dataset expansion. For each sample x_i, we obtain unimodal prediction probabilities p_i^k, $k \in \{a, v\}$. The representational ability of each modality is then assessed by calculating the KL divergence between p_i^k and a uniform distribution as

$$D_{KL}(p_i^k \| U) = \sum_{j=1}^{M} p_{i,j}^k \log\left(\frac{p_{i,j}^k}{u_j}\right), \tag{2}$$

where $p_{i,j}^k$ represents the predicted probability for class j, and $u_j = \frac{1}{M}$ corresponds to the uniform distribution. KL divergence measures the difference between two probability distributions. In the current application scenario, a smaller KL divergence indicates that the output is closer to a uniform distribution, meaning that the separability of the corresponding unimodal output is worse. This can be interpreted as insufficient training for the current modality on this sample. Therefore, we decouple the multimodal inputs for each sample and only retain the modality that performs the worst, masking input from other modalities to achieve accurate unimodal training (Zhang et al., 2024a; Wei et al., 2024). That is, the new training set is as follows:

$$\mathcal{D}' = \left\{ x^{m_*} \mid x \in \mathcal{D}, m_* = \arg\min_{m_* \in K} D_{KL}(m_i) \right\}. \tag{3}$$

By decoupling the multimodal inputs, the strong modality is exposed to fewer samples, while the weak modality is exposed to more, as shown in Figure 3(a). This forces the model to focus more on learning the weak modality during the optimization process and prevents the strong modality from suppressing the weak modality's learning, thereby promoting modality balance, as shown in Figure 3(b), and ensuring more comprehensive multimodal learning. More importantly, our sample-level unimodal evaluation method provides the basis for reassembling unimodal inputs to avoid modality clash.

Reassemble Unimodal Data. After decoupling the multimodal inputs, we address modality laziness but find that

[1]For x_i, if the unimodal predictions are correct with probabilities of 0.9 and 0.6, the Shapley values for both modalities would be the same, but the representational abilities are different.

modality clash remains, as shown in Figure 3(c). For more sufficient multimodal learning, we further analyze the phenomenon and identify that the interference is introduced at the batch level.

Without loss of generality, we assume that the multimodal fusion method selected for the model is Concatenation, as done by others (Fan et al., 2023; Peng et al., 2022; Wei et al., 2024). Let $\boldsymbol{W} \in \mathbb{R}^{M\times(d_{z^a}+d_{z^v})}$ and $\boldsymbol{b} \in \mathbb{R}^M$ represent the parameters of the linear classifier that produces the logits output. We can then express the output of the multimodal model (without softmax) as

$$f(\boldsymbol{x}) = \boldsymbol{W}\left[\phi^a(\theta^a, \boldsymbol{x}^a); \phi^v(\theta^v, \boldsymbol{x}^v)\right] + \boldsymbol{b}. \tag{4}$$

To analyze the optimization process of each modality, we represent $\boldsymbol{W}$ as a block matrix composed of two parts: $[\boldsymbol{W}^a, \boldsymbol{W}^v]$. We can then rewrite Equation 4 as

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{W}^a\left[\phi^a(\theta^a, \boldsymbol{x}^a)\right] + \boldsymbol{W}^v\left[\phi^v(\theta^v, \boldsymbol{x}^v)\right] + \boldsymbol{b} \\ &= \boldsymbol{W}^a\boldsymbol{z}^a + \boldsymbol{W}^v\boldsymbol{z}^v + \boldsymbol{b}. \end{aligned} \tag{5}$$

Assuming that no additional classification heads are added for each modality branch, the loss for each batch is

$$\begin{aligned} \mathcal{L}_{CE} &= -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\left(\frac{e^{f(x_i)_{y_i}}}{\sum_{j=1}^{M} e^{f(x_i)_j}}\right) \\ &= -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\left(\frac{e^{\boldsymbol{W}^a z_i^a + \boldsymbol{W}^v z_i^v + \boldsymbol{b}_{y_i}}}{\sum_{j=1}^{M} e^{f(x_i)_j}}\right). \end{aligned} \tag{6}$$

Observing Equation 6, we see that when using batch gradient descent, even if we decouple the multimodal inputs and mask the dominant modality for each sample, multimodal inputs still coexist within a batch. The inconsistency in optimization directions causes interference between their gradient update directions, ultimately leading to insufficient model training. Therefore, we propose that **cross-modal optimization interference originates at the batch level** and can be addressed by controlling the composition of each batch, referring to this as reassembling unimodal inputs.

We assume that the dataset $\mathcal{D}$ with K modalities can be transformed into $\mathcal{D}'$ through decoupling multimodal inputs. Based on the retained modality, the dataset can be divided into K subsets, denoted as

$$\mathcal{D}^{m_k} = \left\{x_i \mid x_i^j = 0, \forall j \neq k\right\}. \tag{7}$$

The divided subsets satisfy Equation 8, ensuring that our training set does not expand.

$$\begin{aligned} &\bigcup_{k=1}^{K} \mathcal{D}^{m_k} = \mathcal{D}', \\ &\mathcal{D}^{m_i} \cap \mathcal{D}^{m_j} = \emptyset \quad \forall i \neq j \end{aligned} \tag{8}$$

To control the composition of each batch, we reassemble the unimodal inputs based on the decouple assignments and ensure that each batch B_i contains data from only one subset, as shown in Equation 9.

$$B_i \subseteq \mathcal{D}^{m_k}, \quad k \in \{1, 2, \ldots, K\} \tag{9}$$

When sampling from a specific data subset, Equation 6 can be further simplified. For example, when sampling from the data subset corresponding to the video modality, Equation 6 simplifies to

$$\mathcal{L}_{CE} = -\frac{1}{|B|}\sum_{i=1}^{|B|}\log\left(\frac{e^{\boldsymbol{W}^v z_i^v + \boldsymbol{b}_{y_i}}}{\sum_{j=1}^{M} e^{f(x_i)_j}}\right), \tag{10}$$

where the phenomenon of cross-modal optimization interference has been effectively addressed.

In general, the pseudo-code for our method is provided in Algorithm 1. We bridge modality decoupling and reassembling through unimodal evaluation and implement the Data Remixing method to simultaneously tackle modality imbalance and insufficiency.

Algorithm 1 Method of Data Remixing

Input: input data $\mathcal{D} = \{(x_i^1, x_i^2, ..., x_i^K), y_i\}_{i=1,2,...,N}$, number of modalities K, model parameters θ, training epoch E , warm-up epoch E_r
for $e = 0, \cdots, E-1$ **do**
  if $e < E_r$ **then**
    Update model parameters θ with dataset D;
  else
    for $k = 1, \cdots, K$ **do**
      Initialize $\mathcal{D}^{m_k}$: $\mathcal{D}^{m_k} = \emptyset$;
    end for
    for each sample x in D **do**
      Calculate Uni-modal Ability $\{\varphi^1, \varphi^2, ..., \varphi^K\}$ with Equation 2;
      Identify k corresponding to the minimum φ^k;
      Mask $x^j, j \neq k$ to zero;
      Add x into $\mathcal{D}^{m_k}$;
    end for
    for $k = 1, \cdots, K$ **do**
      Update model parameters θ with dataset $\mathcal{D}^{m_k}$;
    end for
  end if
end for

4. Experiments

4.1. Dataset and Experimental settings

CREMA-D (Cao et al., 2014) is an audiovisual dataset for emotion recognition, consisting of 7,442 video clips from

91 actors. The dataset includes six of the most common emotions: *anger*, *disgust*, *fear*, *happy*, *neutral* and *sad*. A total of 2,443 participants rated each clip for emotion and emotional intensity using three modalities: audiovisual, video only, and audio only. The entire dataset is randomly divided into a training and validation set of 6,698 samples and a test set of 744 samples, with a ratio of approximately 9:1.

Kinetic-Sounds (Arandjelovic & Zisserman, 2017) is a dataset derived from the Kinetics dataset (Kay et al., 2017), which includes 400 action classes based on YouTube videos. Kinetic-Sounds consists of 31 action categories, selected for their potential to be represented both visually and aurally, such as playing various instruments. Each video is manually annotated for human actions using Mechanical Turk and is cropped to 10 seconds, focusing on the action itself. The dataset comprises 19k 10-second video clips, split into 15k for training, 1.9k for validation, and 1.9k for testing.

Experimental settings. Unless otherwise specified, all feature extraction networks used in the experiments are ResNet-18 (He et al., 2016), trained from scratch. During training, we use the Adam (Kingma, 2014) optimizer with $\beta = (0.9, 0.999)$ and set the learning rate to 5×10^{-5}. We obtain unimodal prediction results from the unimodal classification heads and also present the results of our method by masking specific modalities to zero, as in Hinton (2012) and Wei et al. (2024). All reported results are averages from three random seeds, with all models trained on two NVIDIA RTX 3090 GPUs using a batch size of 64.

Table 1. Combination and comparison with conventional fusion methods. "+ Remix" indicates that our Data Remixing method is applied. **Bold** indicates that our method brings improvement.

Method	CREMAD	Kinetic-Sounds
Concatenation	64.52%	50.23%
Summation	63.44%	51.66%
Decision fusion	67.47%	52.47%
FiLM	62.77%	49.61%
Bi-Gated	63.04%	49.92%
Concat + Remix	**72.72%**	**55.63%**
Sum + Remix	**71.51%**	**54.09%**
Decision + Remix	**70.70%**	**54.86%**

4.2. Comparison with conventional fusion methods

We first compare our method with several representative multimodal fusion methods commonly used in deep learning frameworks: Concatenation (Concat) (Owens & Efros, 2018), Summation (Sum), Decision Fusion (Decision) (Gunes & Piccardi, 2005), FiLM (Perez et al., 2018), and Bi-Gated (Kiela et al., 2018). The results are shown in Table 1. We apply Data Remixing in combination with Concatenation as the representative fusion method for our approach. It is evident that our method significantly improves model performance across different datasets, with an increase of 8.20% on CREMAD and 5.40% on Kinetic-Sounds, outperforming other conventional fusion strategies. To further demonstrate the generalizability of our method, we combine Data Remixing with Summation and Decision Fusion, achieving substantial improvements across both datasets.

4.3. Comparison with imbalanced multimodal learning methods

Our method is primarily designed to address the issues of modality imbalance and insufficiency in multimodal learning. To evaluate the improvements, we compare our approach with several representative methods that target multimodal balance and sufficiency: G-Blend (Wang et al., 2020), OGM-GE (Peng et al., 2022), Greedy (Wu et al., 2022), PMR (Fan et al., 2023), MLA (Zhang et al., 2024b) and Resample (Wei et al., 2024). For a fair comparison, we adopt Concatenation as the fusion strategy for the baseline, following the approach used in the above methods. The results, including accuracy (Acc) and dataset expansion factor (Factor), are presented for all methods.

Table 2. Comparison with other imbalanced multimodal learning methods. All modulation strategies are applied to the baseline, using Concatenation as the fusion method. We also include results of applying Data Remixing to Resample and MLA for further comparison.

Method	CREMAD		Kinetic-Sounds	
	Acc(%)	Factor	Acc(%)	Factor
Concatenation	64.52	1	50.23	1
+ OGM-GE	68.15	1	52.66	1
+ G-Blend	69.89	1	53.55	1
+ Greedy	68.28	1	51.20	1
+ PMR	68.95	1	51.93	1
+ MLA	68.01	2	54.66	2
+ Resample	67.61	1.85	55.17	1.97
+ Remix (Ours)	**72.72**	**1**	**55.63**	**1**
+ MLA + Remix	**74.19**	1	**57.75**	1
+ Resample + Remix	**73.25**	1.92	**58.40**	2.01

Based on the results in Table 2, we observe that, due to the differing optimization trajectories of modalities, these imbalanced multimodal learning methods show performance improvements over traditional fusion methods. However, our method addresses both modality laziness and modality clash simultaneously, leading to even greater improvements.

Among these, the MLA and Resample strategies also perform relatively well, but hinder training efficiency. Under the same training conditions (especially GPUs and

batch size), MLA employs a multi-stage training strategy and requires gradient modification based on previous features, making its training approximately **3× slower** than our method[2]. The Resample method enhances the model's training on weak modality samples through resampling, which expands the training set by nearly twice, but it hinders training efficiency. To more intuitively reflect the differences in efficiency, we measure the training time of four methods under the same conditions, as shown in Table 3. We observe that balancing methods using unimodal single input tend to increase the training time, whereas Remix does not expand the training set, making it more efficient. Unlike these methods, our approach neither expands the dataset nor reduces the model's training efficiency.

Table 3. Training time to convergence (seconds) on CREMAD and Kinetic-Sounds datasets.

Method	Baseline	Remix	Resample	MLA
CREMAD	1536	2537	4525	6128
Kinetic-Sounds	3849	4946	10362	12868

Considering the similarities between our approach and the Resample and MLA methods—such as MLA's decoupling of unimodal training, which indirectly achieves batch control, and Resample's selective data training—we further combine Data Remixing with these methods to demonstrate its advancements and generalizability. We integrate a sample-level evaluation process into MLA, promoting model balance without increasing the number of input samples (which is why the Factor in Table 2 is updated to 1), and reassemble the unimodal inputs after applying Resample. With our method applied, both approaches show significant improvements in their final results.

4.4. Combination with complex cross-modal architectures

The above methods and experiments are conducted using simple fusion methods, where multimodal fusion occurs after the unimodal encoders or classifiers. To validate the applicability of the Data Remixing method in more complex multimodal scenarios, we combine it with two intermediate fusion methods: MMTM (Joze et al., 2020) and CentralNet (Vielzeuf et al., 2018). MMTM recalibrates the channel features of different CNN streams through squeezing and multimodal excitation steps, while CentralNet uses unimodal hidden representations alongside a central joint representation at each layer, performing fusion through a weighted summation learned during training.

[2]Although MLA does not increase the number of samples, decoupling the inputs without selecting for unimodal training is similar to expanding the dataset, which slows down training.

Table 4. Results on CREMAD and Kinetic-Sounds with two types of complex cross-modal architectures. "+ Remix" indicates the application of our Data Remixing method.

Method	CREMAD	Kinetic-Sounds
Concatenation	64.52%	50.23%
Concat + Remix	**72.72%**	**55.63%**
MMTM	66.40%	52.27%
MMTM + Remix	**68.82%**	**54.78%**
CentralNet	65.46%	54.09%
CentralNet + Remix	**67.61%**	**55.94%**

As shown in Table 4, when multimodal fusion occurs during the encoding process, applying the Data Remixing method leads to significant improvements, further demonstrating the applicability of our method in complex cross-modal architectures.

4.5. Analysis of Methods

In this section, we provide a detailed analysis of the performance improvements introduced by our method. This includes comprehensive ablation experiments and an evaluation of the effectiveness of each design choice. Additionally, we explore different unimodal prediction methods to further demonstrate the broad applicability of the Data Remixing approach.

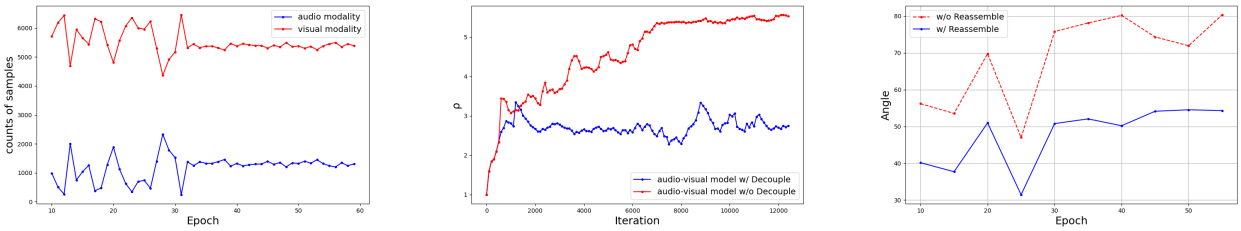
4.5.1. Ablation Study

We conduct ablation studies to demonstrate the effectiveness of our two main designs, with the specific results shown in Table 5. Our experiments consist of two parts: (1) *Decouple*, which involves decoupling the multimodal data and masking a specific modality without reassembling; (2) *Reassemble*, which involves reassembling the modality inputs based on sample-level evaluation without specific training (i.e., without modality masking).

Observing the experimental results, we demonstrate the effectiveness of our method. In Experiment (1), we decouple the multimodal data and select specific samples for training. This approach facilitates modality balance, yielding some performance improvements. In Experiment (2), we reassemble the inputs based on sample-level evaluation, controlling the data composition within each batch. This strategy alleviates modality clash and leads to corresponding improvements. Finally, when both strategies are combined, the model's performance is further enhanced.

4.5.2. Study of Decouple Multimodal Data

In the process of conducting sample-level evaluation and decoupling multimodal data, we choose to mask the better-



(a) (b) (c)

Figure 3. Proof of effectiveness of Decouple and Reassemble Methods. The results are obtained on CREMAD. (a) Statistics of the number of samples used for training specific modalities. (b) The change of the imbalance ratio ρ. (c) Comparison of gradient direction discrepancies.

Table 5. Results of ablation studies on CREMAD and Kinetic-Sounds. The second and third rows correspond to Experiment (1) and Experiment (2), respectively.

Decouple	Reassemble	CREMAD	Kinetic-Sounds
✗	✗	64.52%	50.23%
✓	✗	69.89%	52.31%
✗	✓	68.68%	51.70%
✓	✓	**72.72%**	**55.63%**

performing modality and train with specific samples to alleviate modality imbalance. Figure 3(a) shows the distribution of samples assigned to different modalities during training. It is clear that the strong modality consistently receives fewer samples compared to the weak modality. Figure 3(b) illustrates the change in imbalance ratio ρ (Peng et al., 2022) before and after decoupling the modality inputs. The results indicate that our method alleviates the modality imbalance, supporting our hypothesis that balancing modality representation via sample quantity and difficulty can improve model performance. Note that our method is applied after a 10-epoch warm-up stage.

4.5.3. Study of Reassemble Unimodal Data

To mitigate cross-modal optimization interference, we reassemble unimodal inputs to control the data composition within each batch. Our theoretical analysis of batch control supports the effectiveness of this strategy, and the improvements in unimodal performance are evident. Specifically, audio accuracy on CREMAD increased by 1.75%, while video accuracy improved by 2.96%, since the strong modality is less affected compared to the weak modality.

To further validate the effectiveness of our method in alleviating gradient interference, we evaluate the gradient update discrepancies with and without data reassembling, as shown in Figure 3(c). Specifically, we assess the gradient update discrepancies of the audio modality (the strong modality) on CREMAD at different training stages. We calculate the angle between the actual gradient update direction and the ideal guidance direction for the audio modality. The results show that after reassembling the unimodal inputs, the gradient update direction aligns more closely with the ideal direction, providing further evidence of the effectiveness of our approach.

Table 6. Results of accuracy(%) on CREMAD and Kinetic-Sounds using different methods for unimodal predictions.

Method	CREMAD		Kinetic-Sounds	
	Dropout	Head	Dropout	Head
Concatenation	61.56	64.52	49.58	50.23
Summation	59.68	63.44	48.00	51.66
Decision fusion	62.23	67.47	50.08	52.47
Concat + Remix	**70.03**	**72.72**	**53.89**	**55.63**
Sum + Remix	**68.68**	**71.51**	**53.12**	**54.09**
Decision + Remix	**68.55**	**70.70**	**53.55**	**54.86**

4.5.4. Study of Unimodal Prediction Methods

Since our method requires accurate unimodal predictions, we add a classification head to each modality branch and synchronize the updates of the classification head parameters by modifying the loss function. This approach has been shown to improve model performance (Wang et al., 2020). To further validate our method, we compare the results with the dropout method (Hinton, 2012; Wei et al., 2024) for unimodal prediction evaluation across three traditional fusion strategies. The results are presented in Table 6. From the table, we observe that both methods of obtaining unimodal predictions show effective improvements when applying our strategy. Notably, regardless of the chosen method, our approach does not introduce any additional overhead during inference.

5. Conclusion

Multimodal models often face challenges due to the differences in optimization trajectories between modalities, leading to issues of insufficient and imbalanced learning during joint training. We propose the Data Remixing method, which decouples multimodal data based on unimodal separability and reassembles unimodal data to ensure consistency between modalities at the batch level. Our method effectively mitigates both modality laziness and modality clash, achieving significant performance improvements across various datasets, fusion methods, and model structures. Additionally, it does not require dataset expansion or introduce extra computational overhead during inference. Moreover, to the best of our knowledge, we are the first to propose that the challenges of multimodal learning originate at the batch level and to offer a solution for them. However, a limitation of our approach arises when one modality serves primarily as an auxiliary modality with limited information (Wei et al., 2025). In such cases, the unimodal evaluation and allocation strategy may require further refinement.

Acknowledgments

The work is jointly supported by the National Natural Science Foundation of China (NSFC) under Grant 62261160576 and Grant 62203024, the Beijing Natural Science Foundation (4252026), the Research and Development Program of Beijing Municipal Education Commission (KM202310005027), and the Fundamental Research Funds for the Central Universities of China. This research work is also supported by the Big Data Computing Center of Southeast University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

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

Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., and Zhao, H. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656. PMLR, 2023.

Fan, Y., Xu, W., Wang, H., Wang, J., and Guo, S. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

Fu, J., Gao, J., Bao, B.-K., and Xu, C. Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Ghorbani, A. and Zou, J. Y. Neuron shapley: Discovering the responsible neurons. *Advances in neural information processing systems*, 33:5922–5932, 2020.

Gunes, H. and Piccardi, M. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics*, volume 4, pp. 3437–3443. IEEE, 2005.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, Y., Sun, L., Lian, Z., Liu, B., Tao, J., Wang, M., and Cheng, Y. Multimodal temporal attention in sentiment analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pp. 61–66, 2022.

Hinton, G. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multimodal network fail in deep learning?(provably). In *International conference on machine learning*, pp. 9226–9259. PMLR, 2022.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13289–13299, 2020.

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

Kiela, D., Grave, E., Joulin, A., and Mikolov, T. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kontras, K., Chatzichristos, C., Blaschko, M., and De Vos, M. Improving multimodal learning with multi-loss gradient modulation. *arXiv preprint arXiv:2405.07930*, 2024.

Li, H., Li, X., Hu, P., Lei, Y., Li, C., and Zhou, Y. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22214–22224, 2023.

Lin, B., Lin, Z., Guo, Y., Zhang, Y., Zou, J., and Fan, S. Variational probabilistic fusion network for rgb-t semantic segmentation. *arXiv preprint arXiv:2307.08536*, 2023.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., et al. Multimodal deep learning. In *ICML*, volume 11, pp. 689–696, 2011.

Owens, A. and Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 631–648, 2018.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. 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.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Sun, T., Qian, Z., Li, P., and Zhu, Q. Graph interactive network with adaptive gradient for multi-modal rumor detection. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pp. 316–324, 2023.

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

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12695–12705, 2020.

Wei, Y. and Hu, D. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *International Conference on Machine Learning*, pp. 52559–52572. PMLR, 2024.

Wei, Y., Feng, R., Wang, Z., and Hu, D. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27338–27347, 2024.

Wei, Y., Li, S., Feng, R., and Hu, D. Diagnosing and relearning for balanced multimodal learning. In *European Conference on Computer Vision*, pp. 71–86. Springer, 2025.

Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Xu, R., Feng, R., Zhang, S.-X., and Hu, D. 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)*, pp. 1–5. IEEE, 2023.

Yang, L., Wu, Z., Hong, J., and Long, J. Mcl: A contrastive learning method for multimodal data fusion in violence detection. *IEEE Signal Processing Letters*, 30:408–412, 2022.

Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024a.

Zhang, X., Yoon, J., Bansal, M., and Yao, H. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024b.

Zhou, Y., Liang, X., Zheng, S., Xuan, H., and Kumada, T. Adaptive mask co-optimization for modal dependence in multimodal learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zhu, Y., Wu, Y., Sebe, N., and Yan, Y. Vision+ x: A survey on multimodal learning in the light of data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.