
Mixup Helps Understanding Multimodal Video Better

Xiaoyu Ma^{1 2} Ding Ding^{1 2} Hao Chen^{1 2}

Abstract

Multimodal video understanding plays a crucial role in tasks such as action recognition and emotion classification by combining information from different modalities. However, multimodal models are prone to overfitting strong modalities, which can dominate learning and suppress the contributions of weaker ones. To address this challenge, we first propose Multimodal Mixup (MM), which applies the Mixup strategy at the aggregated multimodal feature level to mitigate overfitting by generating virtual feature-label pairs. While MM effectively improves generalization, it treats all modalities uniformly and does not account for modality imbalance during training. Building on MM, we further introduce Balanced Multimodal Mixup (B-MM), which dynamically adjusts the mixing ratios for each modality based on their relative contributions to the learning objective. Extensive experiments on several datasets demonstrate the effectiveness of our methods in improving generalization and multimodal robustness.

1. Introduction

Multimodal video understanding, as a key research direction in computer vision and multimodal learning (Ngiam et al., 2011), has attracted widespread attention in recent years and plays an important role in tasks such as action recognition (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2019), event detection (Baraldi et al., 2017; Wu et al., 2018), video generation (Clark et al., 2019), and description (Rohrbach et al., 2016). With the development of multimodal learning, various modalities—such as visual, audio, and even textual inputs—have been integrated in the hope of enhancing the model’s representational capacity and task performance through modality complementarity (Zhu et al.,

2024). However, while multimodal inputs provide richer information, they also make models more prone to overfitting specific modalities or spurious correlations in the data. As shown in Figure 1(a), the model quickly converges on the training set, yet achieves relatively low accuracy on the test set, demonstrating data memorization and overfitting.

To alleviate multimodal overfitting and improve generalization, data augmentation techniques have become a major focus of research. Among them, mixup (Zhang et al., 2017) is a simple and effective data augmentation method that generates virtual samples by linearly interpolating between different samples and their labels, thereby enriching the training distribution. Mixup has demonstrated promising results in single-modality tasks such as image classification and speech recognition. However, its exploration in multimodal scenarios such as video understanding remains relatively limited.

To explore the application of mixup in multimodal video understanding, we first proposed the **Multimodal Mixup (MM)** method, which applies mixup at the aggregated multimodal feature level in the feature space to mitigate the model’s tendency to overfit the data (Zhong et al., 2020). This approach of directly applying uniform mixing across modalities is simple and achieves certain benefits. However, it fails to fully account for the dynamic contribution differences (Hu et al., 2022) of each modality during training. Due to the modality imbalance (Wang et al., 2020; Du et al., 2021) in multimodal joint learning—where a strong modality that is easier to optimize can quickly converge and dominate the learning process—the model tends to overfit this strong modality (as illustrated in Figure 1(b)), while other modalities may not have been sufficiently learned. Therefore, applying uniform mixup may not only limit the model’s ability to learn robust cross-modal representations but also exacerbate the risk of overfitting to the strong modality.

To address these issues, we further propose the **Balanced Multimodal Mixup (B-MM)** method for multimodal video understanding. This method monitors the model’s representational capacity on different modalities during training (Peng et al., 2022) and dynamically adjusts the degree of mixing for each modality based on this information. In doing so, it guides the model to learn more balanced mul-

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

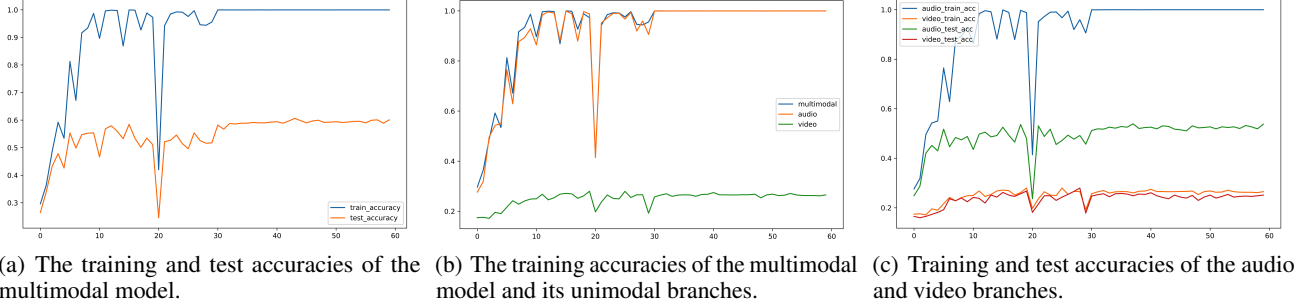


Figure 1. Training and test accuracy curves of the multimodal model and its individual branches during the learning process on the CREMAD dataset.

timodal representations, effectively reducing reliance on any single modality and enhancing generalization in video understanding tasks. We conduct extensive experiments on several benchmark video understanding datasets, and the results demonstrate that our proposed methods consistently improves performance across different tasks, validating the effectiveness of our methods in preventing overfitting and enhancing multimodal cooperation. We summarize our key contributions as follows:

- We introduce a Multimodal Mixup (MM) strategy, which applies the mixup augmentation method to the aggregated multimodal feature space. By creating virtual feature-label pairs, MM enriches the training distribution and effectively mitigates overfitting in multimodal video understanding tasks.
- Building upon MM, we propose the Balanced Multimodal Mixup (B-MM) method to address the modality imbalance issue. B-MM dynamically adjusts mixing ratios for each modality based on their relative contributions during training, promoting balanced representation learning and preventing domination by strong modalities.
- Comprehensive experiments on benchmark datasets (CREMAD, Kinetic-Sounds, and UCF-101) demonstrate that both MM and B-MM consistently outperform traditional fusion and state-of-the-art balanced multimodal learning methods, significantly enhancing model generalization and robustness.

2. Related Work

2.1. Mixup for Data Augmentation

Mixup was originally proposed as a data augmentation method that performs linear interpolation between two samples and their labels (Zhang et al., 2017). It has achieved remarkable generalization performance in single-modality

tasks such as image classification. Building on Mixup, various improved strategies have been proposed in subsequent studies. For example, Manifold Mixup (Cao et al., 2024) performs mixing in the feature embeddings at randomly selected layers rather than only at the raw input level, which helps smooth the decision boundaries and enhance model robustness. MetaMixup (Mai et al., 2021) introduces a meta-learning mechanism that adaptively learns the mixing strategy based on validation performance, alleviating overfitting caused by blind sampling.

In the multimodal domain, extending Mixup to fuse different modalities has led to promising research progress. M3ixup (Lin & Hu, 2024) employs a two-step (adapting and exploring) strategy along with contrastive learning to mix embeddings from different modalities, thereby enhancing the robustness and representational capacity in the presence of missing modalities. Similarly, M3CoL (Kumar et al., 2024) introduces a Mixup-based contrastive loss to better capture cross-modal shared relationships in multimodal contrastive tasks, such as multimodal classification.

2.2. Balanced Multimodal Learning

Existing studies have found that different modalities reach sufficient fitting at different speeds during training (Wang et al., 2020). As a result, when optimizing multimodal models with a unified objective, the strong modality tends to dominate the training process, making the model more prone to overfitting the strong modality while the weak modality remains insufficiently learned (Du et al., 2021).

To mitigate this imbalance, prior work has explored adding additional learning objectives for the weak modality (Wang et al., 2020; Fan et al., 2023; Wei & Hu, 2024; Kontras et al., 2024; Hua et al., 2024) or promoting alignment in optimization rates across modalities (Sun et al., 2021; Peng et al., 2022; Wei et al., 2025; Ma et al., 2025). For example, G-Blending (Wang et al., 2020) calculates the optimal mixing mode of modality losses by determining the overfitting status of each modality. PMR (Fan et al., 2023) introduces a

prototype cross-entropy loss for each modality to accelerate the learning of slower modalities. ATF (Sun et al., 2021) dynamically adjusts the learning rates of different modalities based on the unimodal predictive loss. OGM-GE (Peng et al., 2022) adaptively modulates the gradients of each modality by monitoring discrepancies in their contributions to the learning objective. While these methods promote alignment of optimization rates during training, they often suppress the representational capacity of the strong modality and may be constrained by specific model architectures.

3. Method

3.1. Model Formulation

This work focuses on multimodal video understanding and the modality imbalance phenomenon within it, with downstream tasks including emotion recognition and action recognition. We primarily consider two input modalities: m_a and m_v . The training dataset is denoted as $\mathcal{D} = \{x_i, y_i\}_{i=1,2,\dots,N}$, where each x_i consists of multimodal inputs, i.e., $x_i = (x_i^a, x_i^v)$. The label y_i belongs to $\{1, 2, \dots, M\}$, where M is the number of classes.

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 x . The encoder outputs are represented as $z^a = \phi^a(\theta^a, x^a)$ and $z^v = \phi^v(\theta^v, x^v)$, 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 and denote it as $\mathcal{L}$.

3.2. Mixup in Multimodal Learning

In multimodal supervised learning, we assume that the model receives two modal inputs and obtains the corresponding feature vectors Z^a and Z^v through their respective encoders. Our goal is to identify a function $f \in \mathcal{F}$ that represents the mapping between the feature vectors (Z^a, Z^v) and the target vector Y , where these vectors follow the joint distribution $P((Z^a, Z^v), Y)$. We first define a loss function $\mathcal{L}$ that penalizes the discrepancy between the model’s prediction $f(x^a, x^v)$ and the actual target y for a given example $((x^a, x^v), y) \sim P$. However, since the distribution P is unknown, we approximate the expected risk using the empirical risk computed over the available dataset samples:

$$R_\delta(f) = \int \mathcal{L}(f(x^a, x^v), y) dP_\delta((x^a, x^v), y) \quad (1)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i^a, x_i^v), y_i), \quad (2)$$

where P_δ is the empirical distribution. As noted by Zhang et al. (2017), learning the function f by minimizing Equation 1 can improve computational efficiency. However, when the network has a large number of parameters, a straightforward way to minimize Equation 1 is to memorize the training data, which leads to the well-known issue of overfitting. By observing the multimodal learning process, we find that multimodal models may exhibit data memorization during training. As shown in Figure 1(a), the accuracy of the model on the training set increases rapidly to nearly 100%, while its accuracy on the test set remains significantly lower, indicating poor generalization performance. To improve the model’s performance, we introduce mixup (Zhang et al., 2017) into the field of multimodal learning as shown in Figure 2, where we first get the multimodal feature representations of each sample and then adopt mixup method. Then, virtual feature-target vectors are generated by sampling from a mixed neighborhood distribution:

$$\begin{aligned} \tilde{z}^a &= \lambda \cdot z_i^a + (1 - \lambda) \cdot z_j^a, & \tilde{z}^v &= \lambda \cdot z_i^v + (1 - \lambda) \cdot z_j^v, \\ \tilde{y} &= \lambda \cdot y_i + (1 - \lambda) \cdot y_j, \end{aligned} \quad (3)$$

where $\lambda \sim \text{Beta}(\gamma, \gamma)$, for $\gamma \in (0, +\infty)$. The parameter γ serves as a hyperparameter that controls the strength of data interpolation in mixup. When $\gamma \rightarrow 0$, the model is effectively trained using the conventional Empirical Risk Minimization (ERM) method (Vapnik, 1999).

3.3. Balanced Multimodal-Mixup

Directly and equally applying the mixup method to all modal inputs is a simple and efficient approach. It can effectively mitigate the model’s tendency to memorize training data and improve performance. However, the performance gains achieved by this strategy are quite limited. This phenomenon arises from the issue of modality imbalance in multimodal learning.

When multimodal inputs are jointly trained with a unified objective, the gradients during backpropagation are determined by the combined contributions of all modalities. This can be expressed as follows:

$$\frac{\partial \mathcal{L}}{f(x_i^a, x_i^v)_c} = \frac{e^{(W[z_i^a; z_i^v] + b)}}{\sum_{k=1}^M e^{(W[z_i^a; z_i^v] + b)_k}} - 1_{c=y_i}, \quad (4)$$

where W and b denote the parameters of the final fully connected layer. From Equation 4, we can find that when one modality can be optimized and converge quickly, it tends to dominate the overall learning process, preventing other modalities from being sufficiently learned. As shown in Figure 1(b), during training, the audio modality can learn quickly, whereas the video modality, which contains richer information, fails to be fully learned throughout the process.

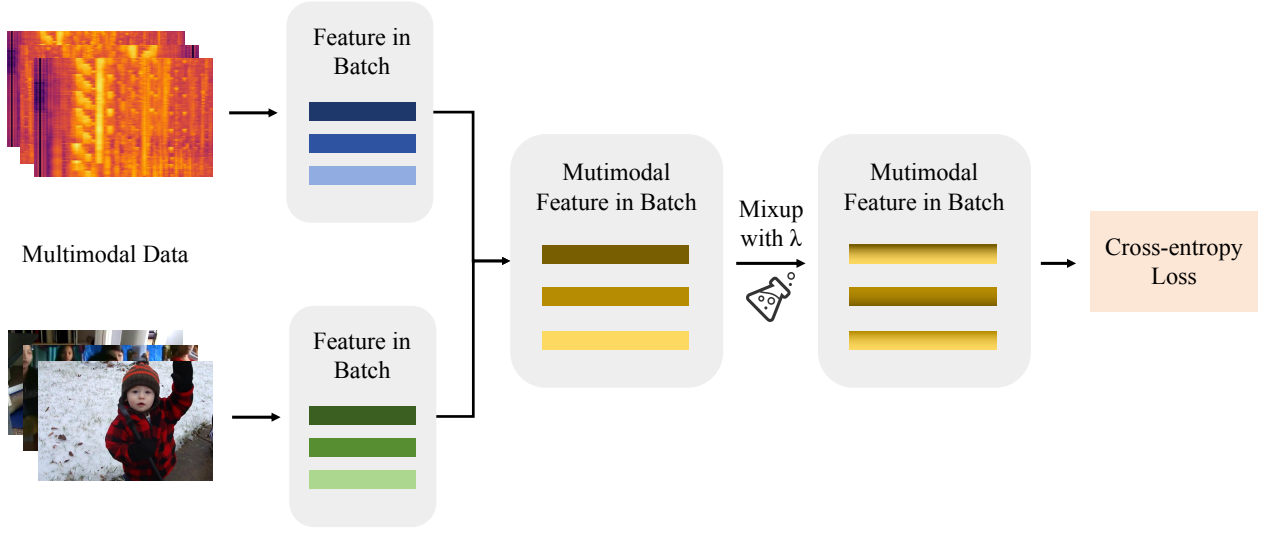


Figure 2. The pipeline of the Multimodal Mixup (MM) method. For each batch, the feature representations of individual modal inputs are first extracted and fused to obtain a multimodal feature representation. The mixup (Zhang et al., 2017) method is then applied to the multimodal feature representation with λ as the mixing parameter to generate virtual feature-label pairs, which are used for model learning.

Moreover, we calculate the train accuracy and test accuracy for both modalities as shown in Figure 1(c) and find that: The audio modality exhibits a much higher accuracy on the training set compared to the test set, reflecting a learning pattern characterized by data memorization. In contrast, the video modality shows only a small difference in accuracy between the training and test sets, indicating that its overall learning remains insufficient.

From this, we observe that in the multimodal learning process, due to the modality imbalance problem, it is often the strong modality that tends to memorize the data and overfit. In such cases, simply mixing multimodal features may not only limit the model’s ability to learn robust cross-modal representations but also exacerbate the risk of overfitting to the strong modality. Therefore, we use the modality differences observed during training as a reference to determine which modality should undergo mixup and to what degree, as illustrated in Figure 3.

Specifically, we first compute the imbalance factor ρ of the

model (Peng et al., 2022) after each training epoch:

$$\begin{aligned} s_i^a &= \sum_{k=1}^M 1_{k=y_i} \cdot \text{softmax}(W[z_i^a; 0] + \frac{b}{2})_k, \\ s_i^v &= \sum_{k=1}^M 1_{k=y_i} \cdot \text{softmax}(W[0; z_i^v] + \frac{b}{2})_k, \end{aligned} \quad (5)$$

$$\rho_e^v = \frac{\sum_{i \in D} s_i^v}{\sum_{i \in D} s_i^a}. \quad (6)$$

Similar to previous work (Peng et al., 2022), we mask one of the modalities to zero and use Equation 5 to obtain the prediction accuracy of each unimodal branch. However, since our method is not applied at every optimization step, we compute the imbalance factor ρ_t^v by aggregating statistics over the entire training set as Equation 6 shows.

By dynamically monitoring the change in ρ in each training round to reflect the contribution differences between the audio and visual modalities, we are able to adaptively adjust the degree of Mixup applied to each modality’s input in the next epoch as follows:

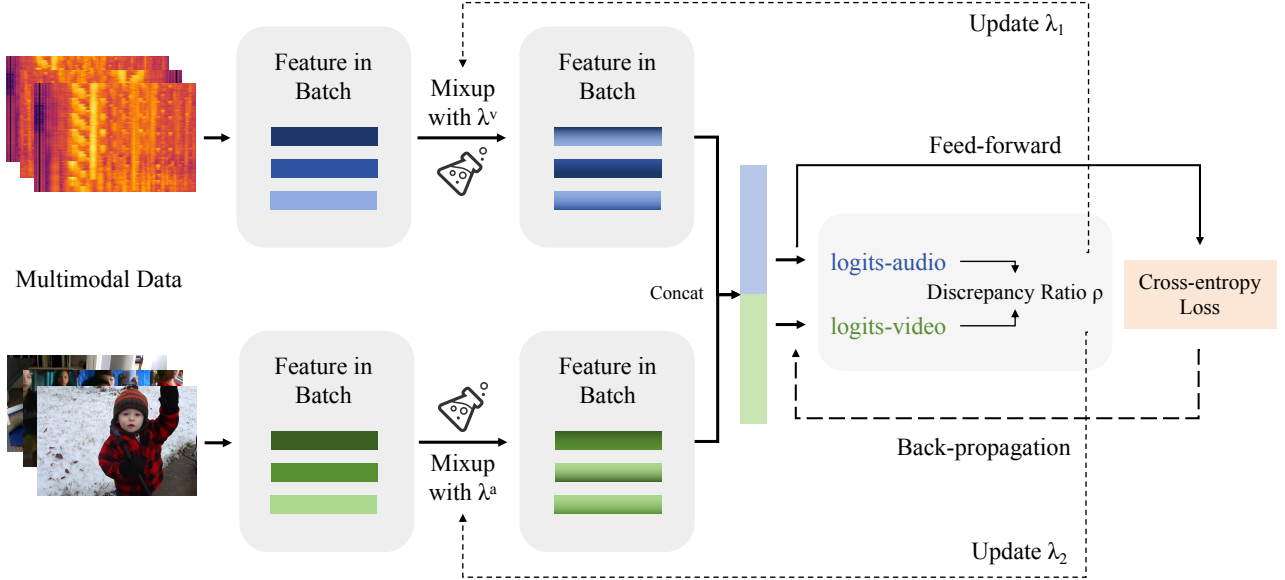


Figure 3. The pipeline of Balanced Multimodal Mixup (B-MM) method. Similar to the MM method, we first get the feature representations of individual modal inputs. Then we apply the mixup method to the unimodal features according to the parameters λ^a and λ^v . After each epoch, the two parameters will update according to the discrepancy ratio ρ (Peng et al., 2022) of modalities.

$$\lambda_t^u = \begin{cases} \tanh(\alpha \cdot \rho_t^u) & \rho_t^u > 1 \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where α is a hyperparameter that controls the degree of mixup. We use the tanh function to regulate ρ , ensuring that it is constrained within the range $[0, 1]$ while preserving monotonicity. Specifically, when the performance of one modality is better, we apply the mixup method to the other modality’s input according to the computed λ_t^u value, as follows:

$$\begin{aligned} \tilde{z}^a &= z_i^a, \\ \tilde{z}^v &= \lambda_t^a \cdot z_i^v + (1 - \lambda_t^a) \cdot z_j^v, \\ \tilde{y} &= \lambda_t^a \cdot y_i + (1 - \lambda_t^a) \cdot y_j, \end{aligned} \quad (8)$$

when $\rho_t^a > 1$, which means audio is the strong modality.

Since the labels are coupled with the modal inputs, this dynamic Mixup applied to the weak modality enables: (1) the strong modality to encounter novel paired modalities and labels, thereby preventing data memorization and overfitting; and (2) the weak modality to generate neighborhood samples, providing greater learning capacity and reducing suppression by the strong modality. This approach helps mitigate modality imbalance during multimodal joint learning and enhances model performance.

4. Experiments

4.1. Dataset and Experimental Settings

This subsection describes the datasets and experimental settings used in the subsequent study. The main experiments in this work are conducted on three video understanding datasets: CREMAD (Cao et al., 2014), Kinetic-Sounds (Arandjelovic & Zisserman, 2017), and UCF101 (Soomro et al., 2012), corresponding to two downstream tasks—emotion recognition and action recognition.

CREMAD is an audio-visual dataset for emotion recognition, consisting of 7,442 video clips performed by 91 actors. The dataset covers six common emotion categories: anger, disgust, fear, happiness, neutral, and sadness. A total of 2,443 raters evaluated the emotion and intensity of each clip using three modalities: audiovisual, video-only, and audio-only. The dataset is randomly split into a training and validation set containing 6,698 samples, and a test set containing 744 samples, with a sample ratio of approximately 9:1 between the training/validation and test sets. In this work, we use video frames and audio as multimodal inputs for video understanding.

Kinetic-Sounds is a dataset derived from the Kinetics (Kay et al., 2017) dataset. The Kinetics dataset contains 400 hu-

Table 1. Combination and comparison with conventional fusion methods. Bold indicates that our method brings improvement, where “+MM” indicates the use of the Multimodal Mixup method, and “+B-MM” indicates the use of the Balanced Multimodal Mixup method. The best results are highlighted in **Bold**, and the second-best results are Underlined.

Method	CREMAD (Acc)	Kinetic-Sounds (Acc)	UCF-101 (Acc)
Concatenation	60.62%	48.50%	79.09 %
Summation	57.80%	48.84%	77.08 %
Decision Fusion	61.83%	49.34%	77.74 %
FiLM	59.68%	48.65%	78.72 %
Bi-Gated	60.89%	49.23%	77.82 %
Concatenation + MM	64.65%	50.89%	80.81 %
Summation + MM	63.58%	50.62%	79.88 %
Decision Fusion + MM	65.86%	51.85%	80.25 %
Concatenation + B-MM	69.22%	53.66%	83.32%
Summation + B-MM	68.15%	52.58%	81.92 %
Decision Fusion + B-MM	<u>68.82%</u>	53.82%	<u>82.47%</u>

man action classes collected from YouTube videos, while Kinetic-Sounds selects 31 action classes that are visually and acoustically distinguishable (e.g., playing musical instruments). Each video is manually annotated for human actions using Mechanical Turk and is trimmed into 10-second clips focusing on the action itself. The dataset includes 14,799 clips for training and 2,594 clips for testing. In this work, we use video frames and audio as multimodal inputs for action recognition.

UCF-101 is a widely used action recognition dataset that provides both RGB frames and optical flow data, enabling multimodal video analysis tasks. The dataset contains 101 classes of human daily activities, with each video sample sourced from real-world YouTube videos. According to the original dataset split, it consists of 9,537 training samples and 3,783 test samples. In this work, we use video frames and precomputed optical flow frames as multimodal inputs for action recognition.

Experimental Settings. The experiments involve three modalities: video, audio, and optical flow. For all video modalities, frames are sampled at 1 fps, and image frames are uniformly selected as inputs. For the audio modality, spectrograms are generated using Librosa (McFee et al., 2015) and used as inputs. Features for all three modalities are extracted using a ResNet-18 (He et al., 2016) network trained from scratch. During training, the Adam (Kingma, 2014) optimizer is used for parameter optimization, with $\beta = (0.9, 0.999)$ and a learning rate set to 5×10^{-5} . All reported results are averaged over three runs with different random seeds, and all models are trained for 60 epochs with a batch size of 64 on two NVIDIA RTX 3090 GPUs to ensure convergence.

4.2. Comparison with Conventional Fusion Methods

To validate the effectiveness of our proposed method, we first combine and compare the proposed Multimodal Mixup (MM) and Balanced Multimodal Mixup (B-MM) methods with several classical multimodal fusion approaches in deep learning. Specifically, these include Concatenation (Concat) (Owens & Efros, 2018), Summation (Sum), Decision Fusion (DeFu) (Gunes & Piccardi, 2005), FiLM (Perez et al., 2018), and Bi-Gated (Kiela et al., 2018). Among them, methods such as concatenation and FiLM belong to mid-level fusion strategies, while decision fusion represents a late fusion strategy. The results are shown in Table 1.

We apply Multimodal Mixup and Balanced Multimodal Mixup in combination with concatenation as the representative fusion method for our approaches. It is evident that Multimodal Mixup serves as an effective data augmentation strategy, significantly improving model performance across different datasets and already outperforming other traditional fusion strategies. Furthermore, by incorporating modality imbalance as a reference factor and designing the Balanced Multimodal Mixup method, the model’s performance is further enhanced, with improvements of 8.60%, 5.16%, and 4.23% on CREMAD, Kinetic-Sounds, and UCF-101, respectively. To further demonstrate the generalizability of our method, we combine data mixing with summation and decision fusion, achieving substantial improvements on two datasets.

To provide a more intuitive comparison of the model’s ability to represent data from each modality before and after applying the BMM method, we perform dimensionality reduction and visualization of the multimodal feature outputs as well as individual unimodal features using UMAP

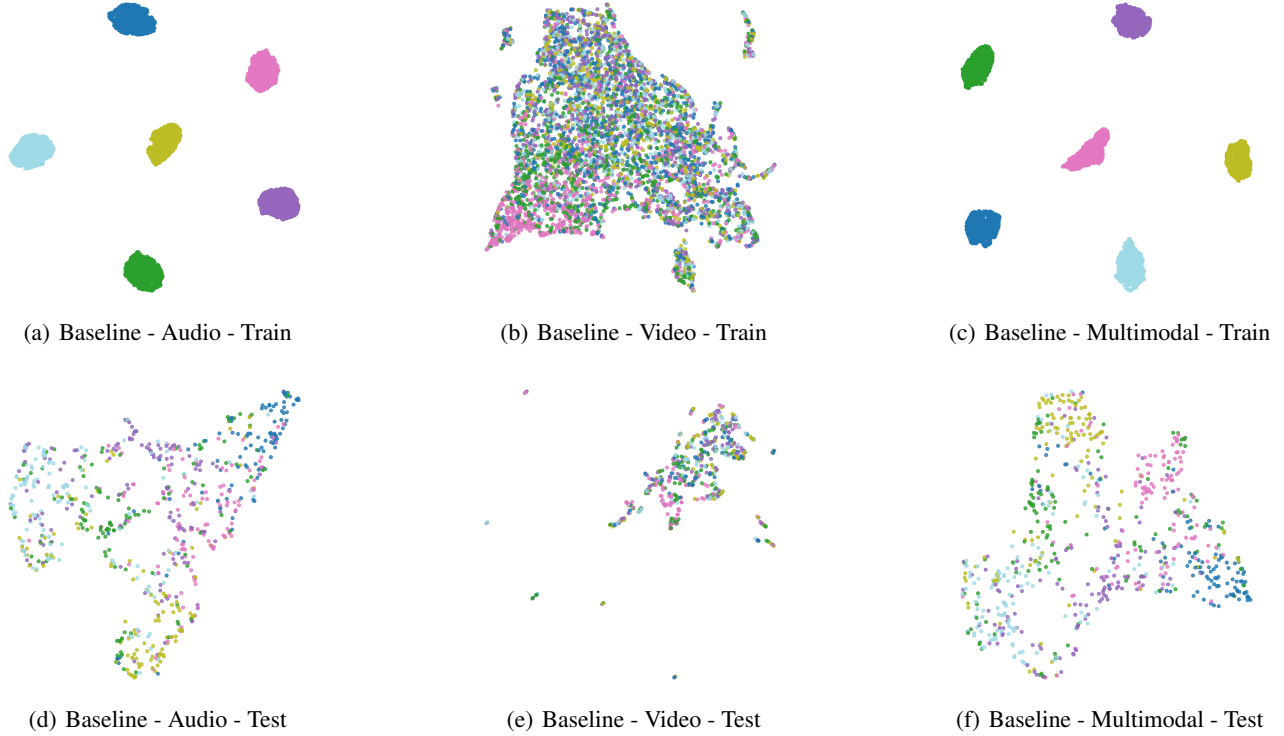


Figure 4. UMAP visualizations of feature representations from the **Baseline** models on the training and test sets. Within each configuration, visualizations for audio, video, and multimodal features are included. Different colors indicate different classes.

Table 2. Comparison with other imbalanced multimodal learning methods. All modulation strategies are applied to the baseline, using Concatenation as the fusion method. The best results are highlighted in **Bold**, and the second-best results are Underlined.

Method	CREMAD	Kinetic-Sounds	UCF-101
Concat	60.62%	48.50%	79.09%
+ G-Blend	67.34%	51.46%	82.21%
+ UMT	65.46%	50.31%	82.10%
+ PMR	67.47%	51.89%	82.34%
+ OGM	<u>68.55%</u>	51.23%	<u>82.55%</u>
+ Greedy	66.13%	<u>52.43%</u>	81.95%
+ ATF	64.25%	51.54%	82.16%
+ MM	64.65%	50.89%	80.81%
+ B-MM	69.22%	53.66%	83.32%

(McInnes et al., 2018), as shown in Figure 4 and Figure 5.

From the visualizations, we observe that when using the baseline multimodal model, the model almost completely memorizes the data from the audio modality (Fig. 4(a)), while the learning performance for the video modality is extremely poor (Fig. 4(b)), with almost no class separability. Furthermore, the final multimodal representations are

essentially dominated by the audio modality (Fig. 4(c)). In contrast, after applying the B-MM method, the performance of the audio modality shows almost no degradation (Fig. 5(a)), but the feature space becomes significantly more compact, consistent with the intended effect of mixup. Meanwhile, the video modality shows a substantial improvement in separability compared to the baseline model (Fig. 5(b)). These results further highlight the importance of B-MM in promoting modality balance during multimodal learning.

4.3. Comparison with Balanced Multimodal Learning Methods

In our mixup process, we take into account the modality imbalance problem in multimodal learning and design the Balanced Multimodal Mixup method based on the discrepancy ratio ρ (Peng et al., 2022). To evaluate the advancement of our approach, we compare it with several representative methods that address multimodal balance and sufficiency, including G-Blend (Wang et al., 2020), UMT (Du et al., 2021), PMR (Fan et al., 2023), OGM-GE (Peng et al., 2022), Greedy (Wu et al., 2022), and ATF (Sun et al., 2021). For fair comparison, we adopt concatenation as the baseline fusion strategy and follow standardized experimental settings commonly used in the field.

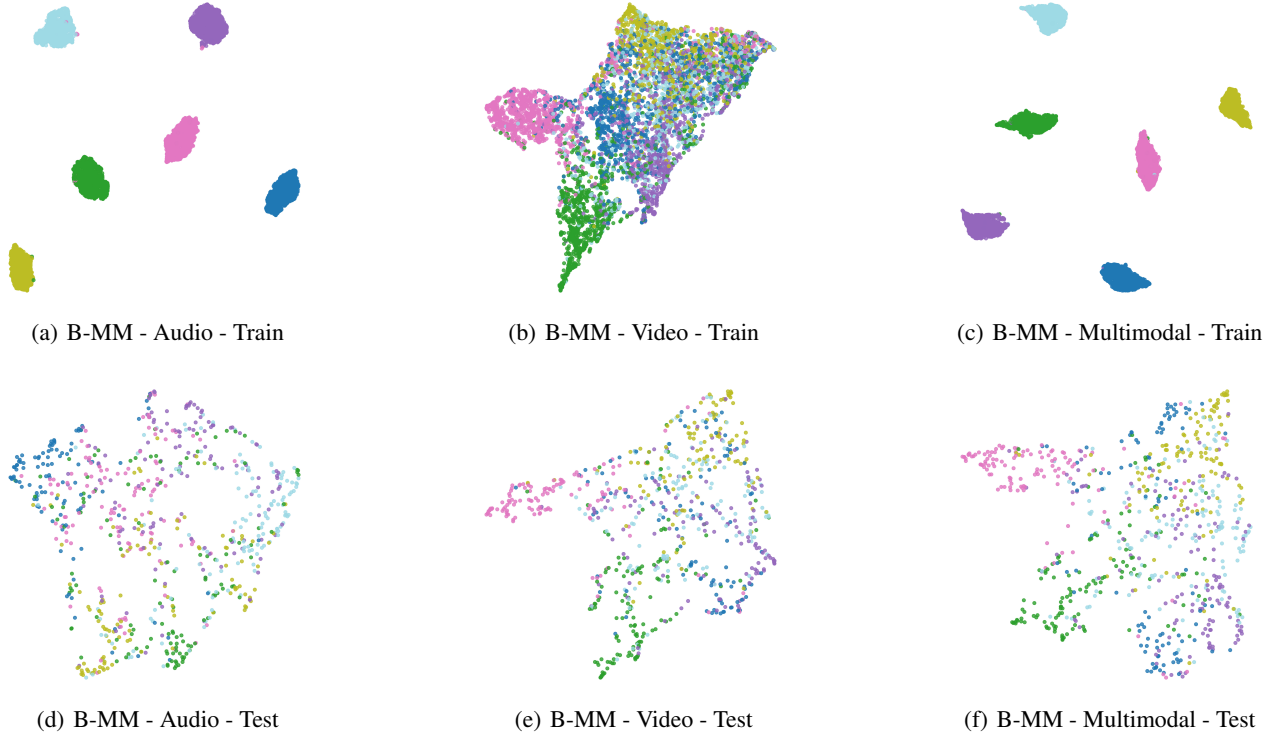


Figure 5. UMAP visualizations of feature representations from the **B-MM** models on the training and test sets. Within each configuration, visualizations for audio, video, and multimodal features are provided. Different colors indicate different classes.

Table 3. Ablation study on the effect of different fixed λ values in the MM method on the CREMAD and Kinetic-Sounds datasets. “(+)” indicates performance improvement over the baseline, while “(-)” indicates performance degradation. The best result for each dataset is highlighted in bold.

Dataset	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CREMAD	61.96% (+)	63.17% (+)	64.65% (+)	63.58% (+)	58.06% (-)	57.39% (-)
Kinetic-Sounds	50.12% (+)	50.89% (+)	49.04% (+)	49.11% (+)	48.46% (-)	47.42% (-)

By examining the results in Table 1 and Table 2, we observe that applying our MM method effectively alleviates model overfitting and improves performance. However, as MM does not address the fact that overfitting mainly arises from the strong modality, its performance still lags behind that of balanced learning methods to some extent. In contrast, the B-MM method achieves greater performance gains, as it accounts for the differences in learning effectiveness across modalities and applies dynamic Mixup accordingly. As a result, B-MM significantly outperforms conventional balance learning approaches.

4.4. Ablation Study

We first conduct an ablation study on the λ parameter in the Multimodal Mixup method to investigate its impact on model performance, as shown in Table 3. We observe that

a very small or large value of λ will result in performance degradation, indicating that excessive or insufficient mixing weakens the benefits of the Mixup strategy.

These results suggest that an appropriate degree of mixing is crucial for balancing data augmentation and preserving the integrity of modal information, and highlight the necessity of adaptively tuning λ for different tasks and datasets.

Next, we conduct ablation studies on two key hyperparameters in the Balanced Multimodal Mixup method: the number of warm-up epochs before applying B-MM and the parameter that controls the degree of Mixup. The results on the CREMAD and Kinetic-Sounds datasets are presented in Table 4 and Table 5, respectively.

By examining the results, we observe that performing an appropriate warm-up phase before applying the B-MM method helps ensure that the model achieves a basic level of perfor-

Table 4. Ablation study on the effect of different numbers (n) of warm-up epochs before applying the B-MM method on the CREMAD and Kinetic-Sounds datasets. The best result for each dataset is highlighted in bold.

Dataset	$n = 0$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$
CREMAD	67.61%	67.61%	69.22%	65.59%	65.46%	64.65%
Kinetic-Sounds	52.70%	53.66%	53.59%	53.20%	52.27%	52.69%

Table 5. Ablation study on the effect of the parameter α that controls the degree of Mixup. The best result for each dataset is highlighted in bold.

Dataset	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
CREMAD	67.47%	69.22%	68.15%	68.28%	67.34%	67.88%
Kinetic-Sounds	51.92%	52.85%	53.66%	52.46%	52.69%	51.65%

mance. This prevents the model from being exposed too early to complex virtual samples, which could otherwise hinder sufficient learning of each modality. In addition, the choice of α should not be too large, as there are inherent interactions between different modalities. A large α value would lead to overly aggressive mixing, preventing the model from adequately learning cross-modal mutual information (Han et al., 2021) and ultimately resulting in performance degradation.

5. Conclusion

In this work, we explored the challenge of modality imbalance in multimodal video understanding and proposed two complementary methods: Multimodal Mixup (MM) and Balanced Multimodal Mixup (B-MM). MM introduces mixup at the multimodal feature level to mitigate overfitting by enriching the training distribution with virtual feature-label pairs. Building on this foundation, B-MM further addresses the imbalance among modalities by dynamically adjusting mixing strategies based on each modality’s contribution during training. Extensive experiments on CREMAD, Kinetic-Sounds, and UCF-101 demonstrated that our methods consistently outperform conventional fusion strategies and existing balanced learning approaches, achieving better generalization and more robust multimodal cooperation.

Despite these promising results, several open questions remain. For instance, how can the dynamic Mixup strategy be extended to handle more than two modalities or to adapt to scenarios with missing or noisy modalities? Furthermore, integrating our approach with large-scale pretrained multimodal models and investigating its impact on tasks beyond classification, such as video captioning or temporal localization, are valuable directions for future research. We hope this work inspires further exploration of adaptive data augmentation for multimodal learning.

References

- Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.
- Baraldi, L., Grana, C., and Cucchiara, R. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1657–1666, 2017.
- Cao, C., Zhou, F., Dai, Y., Wang, J., and Zhang, K. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *ACM Computing Surveys*, 57(2):1–38, 2024.
- 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.
- Clark, A., Donahue, J., and Simonyan, K. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.
- 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.
- 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.
- Feichtenhofer, C., Fan, H., Malik, J., and He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.
- 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.

- Han, W., Chen, H., and Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021.
- 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.
- Hu, P., Li, X., and Zhou, Y. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*, 2022.
- Hua, C., Xu, Q., Bao, S., Yang, Z., and Huang, Q. Re-conboost: Boosting can achieve modality reconciliation. *arXiv preprint arXiv:2405.09321*, 2024.
- 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.
- Kumar, R., Singhal, R., Kulkarni, P., Mehta, D., and Jadhav, K. Harnessing shared relations via multimodal mixup contrastive learning for multimodal classification. *arXiv preprint arXiv:2409.17777*, 2024.
- Lin, R. and Hu, H. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion*, 105: 102216, 2024.
- Ma, X., Chen, H., and Deng, Y. Improving multimodal learning balance and sufficiency through data remixing. *arXiv preprint arXiv:2506.11550*, 2025.
- Mai, Z., Hu, G., Chen, D., Shen, F., and Shen, H. T. Metamixup: Learning adaptive interpolation policy of mixup with metalearning. *IEEE transactions on neural networks and learning systems*, 33(7):3050–3064, 2021.
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24, 2015.
- 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.
- Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 817–834. Springer, 2016.
- Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.
- Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- 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.
- Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- 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., Li, S., Feng, R., and Hu, D. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pp. 71–86. Springer, 2025.
- Wu, C.-Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A. J., and Krähenbühl, P. Compressed video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6026–6035, 2018.
- 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.
- Zhang, H., Cisse, M., Dauphin, Y., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *Learning, Learning*, Oct 2017.
- Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.
- 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.