Understanding Multimodal Learning from Modality Fusion and Alignment Perspectives

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, Tianxi Zhu, and Yi Xu

Abstract—Despite significant advancements in multimodal learning (MML), it has been unexpectedly shown to underperform compared to unimodal approaches in practice, largely due to the modality imbalance problem, ultimately affecting the overall performance of the model. Naturally, most existing methods aim to rebalance optimization speeds across different modalities to avoid performance degeneration caused by modality imbalance. However, in addition to task-oriented modality fusion, we experimentally find that multimodal learning requires *explicit* modality alignment to stimulate weak modal capabilities so that they can be fully exploited, which is ignored by existing works. Therefore, in this paper, we explore the impact of modality fusion and alignment on multimodal learning from a unified perspective, and develops a dynamic strategy that jointly optimizes both, with particular emphasis on addressing modality imbalance. Concretely, we initially design a soft alignment strategy to impose the positive intervention from the prediction level by integrating modality fusion and alignment into a unified framework. We further extend this strategy to the representation level and hybrid level, enabling compatibility with a wider range of architectures. Subsequently, we design a heuristic strategy to dynamically integrate fusion and alignment. Furthermore, we develop a learning-based strategy using a bi-level optimization framework and theoretically prove the convergence of the learning algorithm to ensure its reliability. These two dynamic integration strategies are incorporated into a unified framework applicable to both supervised and semi-supervised scenarios, further enhancing performance. We conduct a series of experiments to demonstrate the effectiveness of our method on diverse datasets. Extensive experiments demonstrate that our method consistently outperforms state-of-the-art multimodal learning approaches, achieving accuracy improvements of 1.30%, 2.69%, and 0.60% on representative bimodal benchmarks, namely KSounds, CREMA-D, and Sarcasm, respectively, as well as gains of 1.35% and 0.65% on trimodal datasets, namely NVGesture and IEMOCAP.

Index Terms—Multimodal Learning, Multimodal Imbalance, Modality Fusion, Modality Alignment.

1 Introduction

In recent years, multimodal learning has experienced explosive growth with the rapid advancement of artificial intelligence. A key challenge in multimodal learning lies in effectively integrating and fully utilizing multimodal information. With advancements in deep learning [1, 2], the representation capacity of multimodal data has significantly improved, thereby enhancing the performance of multimodal models. Thus, multimodal learning has become a hot research topic with a wide range of real applications including image caption [3], cross-modal retrieval [4, 5], vision reasoning [6], action recognition [7], and so on.

By integrating information from multiple modalities, multimodal learning is expected to outperform unimodal models. Unfortunately, several recent studies [8, 9] have revealed an interesting phenomenon, i.e., the performance of the multimodal model is far from the upper bound or even inferior to the unimodal in certain situations. Essentially, the heterogeneity of multimodal data leads to significant differences in the performance of different modalities, with dominant modality being more fully utilized and performing better, while non-dominant modes are less effectively utilized and perform worse [10]. Due to the presence of dominant and non-dominant modalities, the joint training paradigm tends to favor the dominant modality while neglecting the non-dominant modality, which in turn leads to modality imbalance [8]. Therefore, the multimodal model loses its inherent advantage of leveraging multimodal information and may even underperform compared to unimodal models [9, 11, 12, 13]. This notorious phenomenon has been observed across various multimodal task, including emotion recognition [14], sentiment analysis [14], multimodal recommendation [15] and multimodal retrieval [5, 16].

Several recent works [8, 9, 17, 18, 19] have been introduced to balance the impact of dominant and non-dominant modality during training. Early pioneering approaches attempt to modulate key factors in the optimization process that influence the speed of single-modality training, such as learning rate adjustment [20] and gradient modulation [8, 9, 18, 19, 21]. These methods demonstrate that suppressing the optimization of the dominant modality can alleviate the modality imbalance problem to a certain extent. Other representative methods [22, 23, 24, 25, 26] have explored an alternative approach by employing alternating optimization paradigm instead of joint training, while enhancing interactions between different modalities to facilitate multimodal learning. The core idea of this approach is that, within this optimization paradigm, the optimization information from one modality is transferred to the training process of another modality, enabling appropriate adjustments to be made. Besides, several attempts [17, 27, 28], try to introduce extra networks as an auxiliary module to facilitate multimodal learning. By incorporating additional information from external modules, such as a teacher model,

- *Yang Yang, Fengqiang Wan and Qing-Yuan Jiang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.*
  E-mail: {yyang,fqwan,jiangqy}@njust.edu.cn
- *Yi Xu and Tianxi Zhu are with the School of Control Science and Engineering, Dalian University of Technology, Dalian 116081, China.*
  E-mail: {yxu@dlut.edu.cn, zhutianxi3291@mail.dlut.edu.cn}
- *Corresponding author: Yi Xu.*

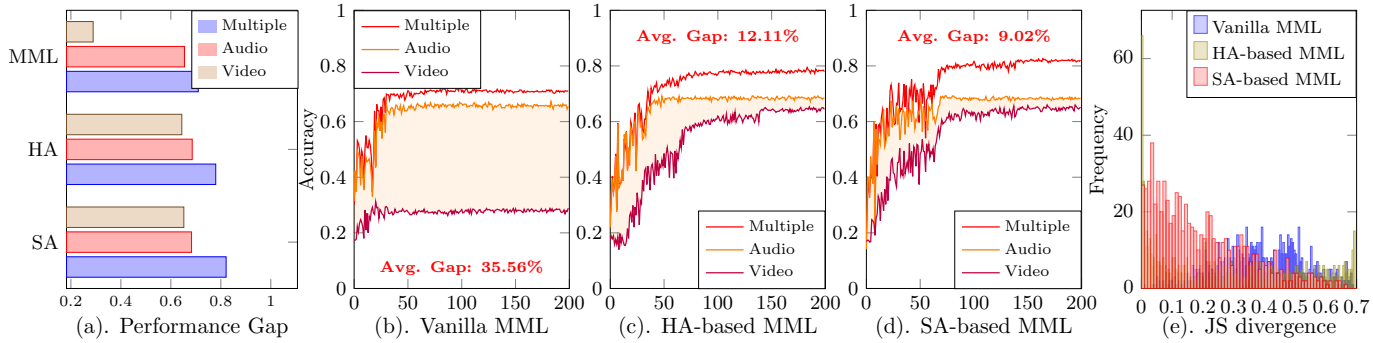


Fig. 1. Illustration of vanilla MML, HA-based MML, and SA-based MML. **(a):** the unimodal performance gap for different methods, where SA-based MML achieves the smallest gap. **(b), (c), (d):** the change of performance gap during training process, showing that alignment-based methods progressively mitigate modality imbalance. **(e):** the histogram of JS divergence for different methods, where SA-based MML yields the lowest divergence. These results indicate that stronger modality alignment leads to a smaller performance gap and more balanced multimodal learning.

the multimodal training of these approaches becomes more balanced.

Almost all aforementioned approaches are based on the phenomenon of inconsistent learning speed itself and do not study the underlying causes of modality imbalance. We can't help but ask what is the essential reason behind this phenomenon? In vanilla multimodal fusion learning, the prediction outcomes of different modalities are expected to align with the ground truth. Ideally, this implies that the predictions from different modalities are implicitly aligned. We take vanilla MML approach with audio and video modalities as an example. The symbols $\mathbf{y}$, $\hat{\mathbf{y}}^a$ and $\hat{\mathbf{y}}^v$ are used to denote the one-hot label, audio prediction and video prediction, respectively. As the fused prediction $f(\hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v)$ converges toward the one-hot label $\mathbf{y}$, the predictions from different modalities are aligned *implicitly*. However, recent studies [8, 9] have revealed that different modalities converge at varying rates, resulting in misaligned predictions that ultimately degrade performance, i.e., modality imbalance. A straightforward enhancement to vanilla MML involves incorporating unimodal loss into the multimodal learning framework, a technique widely adopted in studies such as PMR [18] and MMPareto [29]. In this architecture, unimodal predictions are *explicitly* required to align with the labels, which consequently necessitates their alignment with each other. This procedure can be formed as:

$$\textit{Hard Alignment: } \hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v \rightsquigarrow \mathbf{y} \Rightarrow \hat{\mathbf{y}}^a \leftrightsquigarrow \hat{\mathbf{y}}^v,$$

where "Hard" denotes that the ground-truth label is given in the form of one-hot label. The accuracy in Figure 1 (a) shows that hard alignment (HA)-based MML achieves better performance compared to vanilla MML[1], primarily attributed to the incorporation of alignment within the fusion process. Figure 1 (b) and (c) further reveals that the introduction of hard alignment significantly enhances the performance of weak modalities as training progresses.

Through experimental observations, we find that the lack of *explicit modality alignment* is a key factor contributing to modality imbalance. That is, incorporating modality alignment into task-oriented multimodal learning is essential to mitigate modality imbalance. It is precisely due to the explicit modality alignment that the imbalance caused by fusion is effectively mitigated. Nevertheless, due to the discrete nature of one-hot labels, hard alignment cannot effectively constrain the prediction amplitude. To address this issue, we propose introducing an unsupervised approach [30] to actively intervene in modality fusion. Concretely, we design a soft alignment (SA)-based strategy to impose positive intervention by integrating modality fusion and alignment into a unified framework. The alignment is implemented by using Jensen–Shannon (JS) divergence to minimize the distributional discrepancy across unimodal predictions. This procedure can be formed as:

$$\textit{Soft Alignment: } \hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v \rightsquigarrow \bar{\mathbf{y}} \Rightarrow \hat{\mathbf{y}}^a \leftrightsquigarrow \hat{\mathbf{y}}^v,$$

where $\bar{\mathbf{y}} \triangleq \frac{1}{2}(\hat{\mathbf{y}}^a + \hat{\mathbf{y}}^v)$. We further illustrate the training details in Figure 1, where we utilize SA-based MML to denote the method with fusion and alignment integration strategy. We observe that the unimodal performance gap aligns remarkably well with the degree of alignment in terms of JS divergence. This implies that integrating fusion and alignment from a unified perspective can effectively balance the prediction capability of each modality, thereby enhancing the model performance. To better incorporate alignment into modality fusion, we initially focus on modality alignment and gradually incorporate task-oriented modality fusion during training. Moreover, since prediction-level alignment is not well-suited for some complex fusion structures, we extend this concept to the representation level and hybrid level. Then, we design a simple yet effective heuristic strategy and a learning-based strategy based on bi-level optimization to dynamically integrating fusion and alignment, where the convergence of the optimization algorithm under the learning-based strategy is theoretically proved to ensure its reliability. This unified integration strategy is applicable to both supervised and semi-supervised multimodal learning scenarios. Accordingly, we design specific objective formulations tailored to these two settings. Our contributions are listed as follows:

- We observed a striking consistency between modality performance gap and alignment, which prompted us to explore the phenomenon of modality imbalance from the perspectives of modality fusion and alignment.

1. We provide the implementation details in the appendix.

- Based on our observations, we propose a novel prediction-level soft alignment strategy that integrates modality fusion and alignment into a unified framework, and further extend this concept to the representation level and hybrid level. Since the importance of modality alignment varies during training, with early-stage alignment reducing cross-modal discrepancy and later-stage fusion improving task-oriented prediction, we introduce a simple yet effective heuristic strategy and a learning-based strategy to dynamically coordinate modality alignment and modality fusion. The convergence of the learning-based optimization algorithm is theoretically proved to ensure its reliability.
- We conduct comprehensive experiments under both supervised and semi-supervised scenarios. Experimental results demonstrate that our method can outperform all baselines to achieve SOTA performance.

The remainder of this paper is organized as follows. Section 2 reviews related work on rebalanced multimodal learning and modality alignment. Section 3 introduces the proposed unified framework that integrates modality fusion and alignment, including prediction-level and representation-level alignment strategies, as well as heuristic and learning-based dynamic integration methods with theoretical convergence analysis. Section 4 presents extensive experimental results under both supervised and semi-supervised settings. Finally, Section 5 concludes the paper and discusses future research directions.

2 Related Work

2.1 Rebalanced Multimodal Learning

Multimodal machine learning [1, 2] serves as a bridge, enabling humans to perceive the world more comprehensively and deeply. Given the heterogeneity of data, a key challenge in multimodal machine learning is effectively integrating information from diverse modalities. Based on fusion strategy, multimodal learning can be categorized into early fusion [8, 31], late fusion [32], and hybrid fusion [33]. By integrating information from multiple modalities, multimodal learning is expected to outperform unimodal models.

Unfortunately, recent works [8, 9, 18, 19] have highlighted that modality imbalance is a pervasive challenge, often resulting in suboptimal performance or even worse outcomes compared to unimodal algorithms in certain cases. Given the existence of dominant and non-dominant modalities, early approaches [8, 9, 18] focus on adjusting the learning rates for different modalities through tailored strategies to balance the optimization of both dominant and non-dominant modalities. For example, G-Blend [8] minimizes the overfitting-to-generalization ratio (OGR) using a gradient blending technique based on the overfitting behavior of each modality. OGM [9] applies an on-the-fly gradient modulation strategy to control the optimization process of each modality. PMR [18], on the other hand, introduces a prototypical modality rebalance strategy to facilitate the learning of non-dominant modalities. Other approaches [17, 27] have explored the use of auxiliary networks to address modality imbalance. Specifically, UMT [17] employs teacher networks to distill pretrained unimodal features into the multimodal network, helping to mitigate modality imbalance. Balanced multimodal learning [27] defines conditional learning speeds using the gradient norm and the norms of model parameters, guiding the learning procedure to ensure balance. While all of the above methods belong to the joint training paradigm, more recent research has introduced an alternative training paradigm [22, 23, 24]. Specifically, [22] proposes an alternating training algorithm that iteratively updates different modalities, promoting stronger interactions through enhanced information transfer between modalities. In parallel, [24] incorporates gradient boosting to balance modality learning, leveraging cross-modal information during the interactive training phase. These recent methods mitigate the modality imbalance problem by fostering enhanced interaction between modalities and ensuring more balanced learning. Despite their empirical success, existing rebalanced multimodal learning methods largely share a common assumption: modality imbalance can be alleviated by regulating optimization dynamics alone. As a result, modality interactions are implicitly governed by task loss, without explicitly enforcing semantic or representational alignment during fusion. This often allows dominant modalities to dominate feature aggregation, limiting the effective exploitation of complementary information. In contrast, we study modality alignment as an important component of fusion and propose a unified framework that jointly integrates fusion and alignment to alleviate modality imbalance from a complementary perspective.

2.2 Modality Alignment

Modality alignment has been a critical area of research in multimodal learning, aiming to align features across different modalities (e.g., text, image, audio) to enable more coherent and meaningful interactions. Based on the alignment strategy, methods can be categorized into prediction-level [34] and representation-level [35, 36] alignment.

Prediction-level alignment targets aligning the predictions from different modalities, minimizing cross-modal differences and enhancing the extraction of intrinsic inter-modal correspondences. For example, [34] uses Kullback–Leibler (KL) divergence to align the predictions of different modalities by minimizing the discrepancy between their predicted distributions, reducing inter-modal inconsistencies. Furthermore, related research for representation-level alignment initially began with canonical correlation analysis (CCA) [35], which aims to explore and quantify the relationships between different modalities. As deep learning gained prominence, deep learning techniques are employed to uncover the underlying relationships among multiple modalities. Deep canonical correlation analysis (DCCA) [36] is an advanced extension of CCA that leverages deep neural networks to model complex, non-linear relationships between two modalities. Furthermore, researchers began to explore joint embedding spaces where features from multiple modalities are aligned through embedding layers, e.g., VILBERT [37]. Contrastive learning methods such as CLIP [38] further revolutionized modality alignment.

3 Methodology

In this section, we present the proposed approach for ad-

TABLE 1
Advantages and limitations of representative rebalanced multimodal learning methods. Existing approaches alleviate modality imbalance mainly by regulating optimization dynamics (e.g., gradient reweighting, auxiliary supervision, or alternating training). However, they typically lack explicit cross-modal guidance to enhance the learning of non-dominant modalities.

Method Category	Main Advantage	Main Limitation
Gradient reweighting (G-Blend [8], OGM [9], PMR [18], AMSS [21])	Explicitly regulates gradient contributions to stabilize optimization	Does not explicitly leverage cross-modal guidance to facilitate learning of non-dominant modalities.
Auxiliary network-based (UMT [17])	Provides additional supervision to enhance learning of weaker modalities	
Alternating training (MLA [22], ReconBoost [24], DUS [25], AUG [26])	Encourages balanced modality updates through interactive optimization	

dressing modality imbalance. As illustrated in Figure 2, our approach comprises two key components: modality fusion module and modality alignment module. The modality fusion module is built upon a standard multimodal fusion architecture, with details provided in Section 3.1. And the modality alignment module is implemented at both the prediction level and the representation level, aiming to enhance cross-modal consistency, as described in Section 3.2. In addition, we introduce dynamic integration strategies to balance the task-oriented fusion objective and the alignment objective during training. We then elaborate the implementation of the proposed method in both supervised and unsupervised multimodal learning, followed by a theoretical analysis of the algorithm.

3.1 Modality Fusion

For the sake of simplicity, we leverage audio and video modalities to illustrate our method. Please note that our method can be easily adapted to cases with more than two modalities. Suppose that we are given n data points, each of which contains audio and video modalities. We use $\mathbf{X} = \{\mathbf{x}_i = (\mathbf{x}_i^a, \mathbf{x}_i^v)\}_{i=1}^n$ to denote the multimodal data point set, where $\mathbf{x}_i^a$ and $\mathbf{x}_i^v$ respectively denote the i-th data point of audio and video modalities. In practice, the category labels of same points are available. We utilize $\mathbf{Y} = \{\mathbf{y}_i \mid \mathbf{y}_i \in \{0,1\}^c\}_{i=1}^m$, where c and m respectively denote the number of category labels and the number of data points with labels being available. If $m = n$, the scenario corresponds to standard supervised learning; otherwise, it falls under semi-supervised learning. $\forall y_{ij} \in \mathbf{y}_i$, $y_{ij} = 1$ if $\mathbf{x}_i$ belongs to j-th category, and $y_{ij} = 0$ otherwise. With the advancement of deep learning, multimodal methods leveraging deep learning techniques have yielded significant results. In this paper, we also adopt deep neural network to perform multimodal feature learning following the setting of [8, 10, 24].

We first extract unimodal features for each modality. Concretely, we use $\phi^o(\cdot)$ to denote the feature extraction function, i.e., encoder network, for a specific modality, where $o \in \{a, v\}$. The feature extraction process can be formulated as:

$$\mathbf{z}_i^o = \phi^o(\mathbf{x}_i^o; \Phi^o), \tag{1}$$

where $\mathbf{z}_i^o \in \mathbb{R}^d$ and Φ^o respectively denote d-dimensional features and the parameters of the encoder network.

These unimodal representations are then fused for task prediction. After extracting vectors for all modalities, we apply a fusion function $f(\cdot)$ to combine the different feature vectors. Then, we leverage a multi-layer network to map the vector into $\mathbb{R}^c$. The fusion procedure can be formed as:

$$\mathbf{z}_i = f(\mathbf{z}_i^a, \mathbf{z}_i^v); \quad \hat{\mathbf{y}}_i = \texttt{softmax}(g(\mathbf{z}_i)). \tag{2}$$

Here, $g(\cdot)$ denotes the classification network. Eq. (2) defines the task-oriented multimodal prediction process. Following most existing methods [22], we adopt the concatenation function as the fusion method and a one layer network as classifier for illustration. Then, the objective function of modality fusion can be formed as:

$$L_{\text{Fusion}}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum \ell(\hat{\mathbf{y}}, \mathbf{y}), \tag{3}$$

where the form of $\ell(\cdot)$ depends on the specific task. This objective function leverages supervision to guide model learning and can be instantiated for specific tasks (e.g., classification).

3.2 Modality Alignment

The modality alignment process aims to build the consistency across different modalities. We consider modality alignment at both the prediction level and the representation level. Generally speaking, distribution metrics, such as JS divergence, are widely used tools for measuring differences between distributions. As JS divergence is a symmetrized and smoothed version of KL divergence, we leverage the distribution metric to measure the alignment of prediction across modalities.

For prediction-level alignment, we encourage different modalities to produce consistent predictive distributions. Concretely, we define the logits of audio and video as: $\hat{\mathbf{y}}^a$, $\hat{\mathbf{y}}^v$. We define the fusion function in Eq. (2) with concatenation function and one layer classifier:

$$f(\mathbf{z}_i^a, \mathbf{z}_i^v) = \text{CONCAT}(\mathbf{z}_i^a, \mathbf{z}_i^v) = [\mathbf{z}_i^a, \mathbf{z}_i^v]. \tag{4}$$

Then, we redefine the linear parameter $\mathbf{W}$ as: $\mathbf{W} \triangleq [\mathbf{W}^a, \mathbf{W}^v]$, $\mathbf{W}^a, \mathbf{W}^v \in \mathbb{R}^{c \times d}$. By applying a linear projection

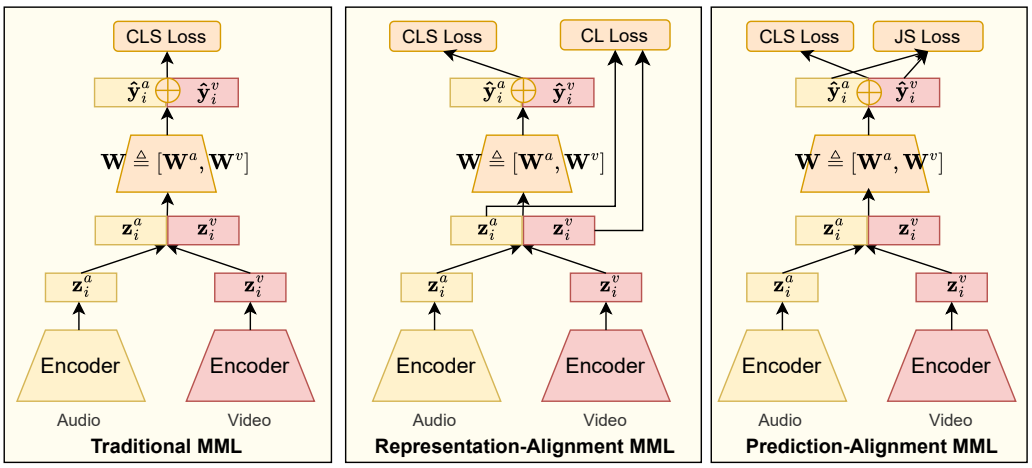


Fig. 2. Illustration of traditional MML and alignment-enhanced MML with representation-level and prediction-level alignment. Traditional MML performs modality-specific feature learning followed by multimodal fusion. The proposed framework further introduces modality alignment to enforce cross-modal consistency, which is implemented at both the representation level and the prediction level during training.

to the fused features followed by a softmax function, the prediction $\hat{\mathbf{y}}^o$ can be obtained:

$$\forall o \in \{a, v\}, \hat{\mathbf{h}}_i^o = \mathbf{W}^o \mathbf{z}_i^o + \frac{\mathbf{b}}{2}, \tag{5}$$

$$\hat{\mathbf{y}}^o = \texttt{softmax}(\hat{\mathbf{h}}^o). \tag{6}$$

The JS divergence of $\hat{\mathbf{y}}^a$ and $\hat{\mathbf{y}}^v$ can be formed as:

$$L_{\text{JS}}(\hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v) = \frac{1}{2}\mathcal{D}_{\text{KL}}(\hat{\mathbf{y}}^a \parallel \bar{\mathbf{y}}) + \frac{1}{2}\mathcal{D}_{\text{KL}}(\hat{\mathbf{y}}^v \parallel \bar{\mathbf{y}}). \tag{7}$$

Here $\mathcal{D}_{\text{KL}}(\cdot)$ denotes the KL divergence. Eq. (7) penalizes the discrepancy between unimodal predictions and thus enforces cross-modal consistency at the prediction level.

We further impose representation-level alignment to improve the consistency of unimodal feature representations. At representation level, we utilize contrastive learning to perform modality alignment. Specifically, the cosine similarity between features $\mathbf{z}_i^a$ and $\mathbf{z}_i^v$ can be formulated as:

$$s(\mathbf{x}_i^a, \mathbf{x}_j^v) = \langle \frac{\mathbf{z}_i^a}{\|\mathbf{z}_i^a\|}, \frac{\mathbf{z}_j^v}{\|\mathbf{z}_j^v\|} \rangle, \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. The objective of contrastive learning can be formed as:

$$L_{\text{CL}}(\mathbf{X}) = -\frac{1}{2n_b}\sum_{i=1}^{n_b} \log\left(\frac{\exp(s(\mathbf{x}_i^a, \mathbf{x}_i^v)/\tau)}{\sum_{k=1}^{n_b}\exp(s(\mathbf{x}_i^a, \mathbf{x}_k^v)/\tau)}\right) \tag{9}$$

$$-\frac{1}{2n_b}\sum_{i=1}^{n_b} \log\left(\frac{\exp(s(\mathbf{x}_i^a, \mathbf{x}_i^v)/\tau)}{\sum_{k=1}^{n_b}\exp(s(\mathbf{x}_k^a, \mathbf{x}_i^v)/\tau)}\right), \tag{10}$$

where n_b, τ denote the batch size and temperature, respectively. Eq. (10) pulls paired cross-modal representations closer while pushing unmatched pairs apart, thereby improving representation consistency across modalities.

Based on the above formulations, a unified alignment objective is defined to accommodate different alignment strategies:

$$L_{\text{Align}} = \begin{cases} L_{\text{JS}}(\hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v), \text{prediction-level alignment}, \\ L_{\text{CL}}(\mathbf{X}), \text{representation-level alignment}, \\ L_{\text{JS}}(\hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v) + L_{\text{CL}}(\mathbf{X}), \text{hybrid alignment}, \end{cases} \tag{11}$$

where L_{JS} and L_{CL} are defined in Eqs. (7) and (10), respectively.

Discussion: During the process of modality alignment, is it possible for the dominant modality to degrade? We observe that the introduction of modality alignment did indeed affect the performance of the strong modality in the initial stages experimentally. However, the overall performance improved in the later stages, indicating that, on the whole, the impact of alignment is positive. Furthermore, we employ a single-layer classifier to illustrate the details of modality alignment, following existing methods. For more complex architectures with additional layers, the alignment can be achieved using a design incorporating unimodal losses. Relevant details are provided in the appendix.

3.3 Integrating Fusion and Alignment

Thus far, we discuss the task-oriented modality fusion and prediction- and representation-level alignment. In this section, we design both heuristic and learning-based integration strategies to construct a unified objective function for learning. Concretely, by integrating modality fusion and alignment, the unified objective function can be obtained by:

$$L_{\text{Total}}(\mathcal{D}) = (1-\alpha)L_{\text{Fusion}}(\mathcal{D}) + \alpha L_{\text{Align}}(\mathcal{D}). \tag{12}$$

Here, a fixed coefficient α is used to integrate the fusion and alignment objectives into a unified objective function. However, such a formulation fails to dynamically balance the contributions of the two objectives throughout training. In essence, our goal is to treat modality alignment as a form of regularization on feature learning in multimodal settings.

This constraint is expected to be stronger in the early stages of training, and gradually diminish as the model shifts its focus toward multimodal fusion. Therefore, α is expected to take a relatively large value in the early training stage and gradually decrease over time. This can be realized either by heuristically defining α as a function of the training iterations, or by directly learning α. We present both the heuristic and learning-based strategies in the following.

Heuristic Integration: Firstly, we present the heuristic integration strategy. To ensure features are sufficiently aligned before performing task-oriented fusion, we prioritize modality alignment by employing a monotonically decreasing function of the number of iterations to determine the value of α. Concretely, we utilize a monotonically decreasing function to obtain α_t at t-th iteration:

$$\alpha_t = 1 - \exp\left(-\frac{1}{t}\right). \tag{13}$$

Based on the above formulation, α is defined as an exponentially decaying function of the training iteration. As t increases, the value of α gradually decreases.

Learning-based Integration: Then, we further exploit a learning-based integration method by utilizing bi-level optimization strategy [39]. Specifically, while considering optimizing the modality fusion loss L_{Fusion}, we use the minimum value of the total loss L_{Total} to restrict the feasible region of the parameters θ. In other words, the parameters are optimized under the guidance of both task-oriented fusion and cross-modal alignment, while α is learned to determine how much alignment should be imposed at different training stages. The corresponding optimization problem is defined as follows:

$$\begin{aligned} &\min_{0\le\alpha\le 1} L_{\text{Fusion}}(\Theta^*(\alpha)) \\ &\text{subject to: } \Theta^*(\alpha) \in \operatorname{argmin}_{\Theta}\ L_{\text{Total}}(\mathcal{D}^{(\text{train})}), \end{aligned} \tag{14}$$

where, Θ denotes the model parameters, and α is the integration parameter that controls the relative importance of modality fusion and modality alignment. Instead of manually fixing this parameter, the learning-based strategy updates α according to the optimization status of the model, so that alignment is emphasized when modality discrepancy is large and fusion becomes more important when the modalities have been sufficiently coordinated. The optimal parameter set, $\Theta^*(\alpha)$, thus represents a fine-tuned balance that, for any chosen α, strategically minimizes this composite loss. Within Eq. (14), L_{Fusion} is pivotal for guiding classification accuracy, while L_{Align} enhances the model's ability to establish meaningful connections across different modalities.

We utilize an approximation method proposed by [40] to solve bi-level optimization problem (14) efficiently. Specifically, the gradient of $L_{\text{Fusion}}(\theta(\alpha))$ with respect to α at the k-th iteration is approximated as follows:

$$\begin{aligned} &\widetilde{\nabla} L_{\text{Fusion}}(\alpha_k, \widehat{\Theta}_k, \varrho_k) = \\ &\qquad - \widetilde{\mathcal{M}}(\alpha_k, \widehat{\Theta}_k, \zeta_k^{(2)}, \zeta_k^{(3)}) \widetilde{\nabla}_{\Theta} G_{\text{Fusion}}(\widehat{\Theta}_k, \zeta_k), \end{aligned} \tag{15}$$

where $\varrho_k = \left(\zeta_k, \zeta_{[k]}^{(1)}, \zeta_k^{(2)}, \zeta_k^{(3)}\right)$, $\widetilde{\nabla}_{\Theta} G_{\text{Fusion}}(\widehat{\Theta}_k, \zeta_k)$ denotes the stochastic vector obtained by accessing the stochastic gradient oracle of $L_{\text{Fusion}}(\widehat{\Theta}_k)$, which is used to approximate $\nabla L_{\text{Fusion}}(\widehat{\Theta}_k)$ with historically independent random variable ζ_k. And the term $\widetilde{\mathcal{M}}(\alpha_k, \widehat{\Theta}_k, \zeta_k^{(2)}, \zeta_k^{(3)})$ in Eq. (15) is defined as follows:

$$\begin{aligned} &\widetilde{\mathcal{M}}(\alpha_k, \widehat{\Theta}_k, \zeta_k^{(2)}, \zeta_k^{(3)}) \triangleq \\ &\qquad \widetilde{\nabla}_{\alpha\Theta} G_{\text{Total}}(\alpha_k, \widehat{\Theta}_k, \zeta_k^{(2)}) \mathcal{H}_{\Theta\Theta}\left(\alpha_k, \widetilde{\Theta}, \zeta_k^{(3)}\right), \end{aligned} \tag{16}$$

where $\widetilde{\nabla}_{\alpha\Theta} G_{\text{Total}}(\alpha_k, \widehat{\Theta}_k, \zeta_k^{(2)})$ is formulated in a manner similar to $\widetilde{\nabla}_{\Theta} G_{\text{Fusion}}(\widehat{\Theta}_k, \zeta_k)$, and $\mathcal{H}_{\Theta\Theta}\left(\alpha_k, \widetilde{\Theta}, \zeta_k^{(3)}\right)$ is the stochastic hessian inverse approximation of $[\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)]^{-1}$. Based on the approximation, we can use the gradient descent method to optimize α.

Algorithm 1: The proposed algorithm.

Input : Training set $\mathcal{D}$.
Output: Learned parameters $\{\Theta\}$ of all models.

INIT initialize parameters Θ, parameter α, maximum iterations T, learning rate $\{\eta_k^{\text{in}}\}_{k\ge 0}$, $\{\eta_k^{\text{out}}\}_{k\ge 0}$.
for $k = 1$ **to** T **do**
  /* updating neural network parameters Θ. */
  for $t = 1$ **to** *Inner_Iters* **do**
    Randomly construct a mini-batch $\mathcal{D}^{(\text{batch})}$.
    Calculate total loss L_{Total} by forward phase.
    Calculate gradient based on SGD.
    Update parameters Θ according to its gradient:

$$\Theta_{t+1} = \Theta_t - \eta_t^{\text{in}} \widetilde{\nabla}_{\Theta} G_{\text{Total}}(\alpha_k, \Theta_k, \zeta_t^{(1)}).$$

  end
  /* updating weighting parameters α. */
  Set $\widehat{\Theta}_k = \Theta_k$.
  Calculate loss L_{Total} in $\mathcal{D}^{(\text{batch})}$ by forward phase.
  Calculate the stochastic gradient approximation of L_{Fusion} according to Eq. (15).
  Calculate α_{k+1} by:

$$\mathcal{P}_k(x) \triangleq \widetilde{\nabla} L_{\text{Fusion}}(\alpha_k, \widehat{\Theta}_k, \varrho_k) \cdot x + \frac{1}{2\eta_k^{\text{out}}} |x - \alpha_k|^2,$$

$$\alpha_{k+1} = \operatorname*{argmin}_{x\in[0,1]} \mathcal{P}_k(x)$$

end

3.4 Supervised and Semi-Supervised MML

The unified architecture in Eq. (12) is adaptable to a variety of application scenarios. Here, we first detail its adaptation to supervised learning scenarios.

If the category label is available, we take the cross entropy loss for classification loss as an example:

$$\ell_{\text{Labeled}}(\mathbf{x}_i, \mathbf{y}_i) = -\mathbf{y}_i^{\top} \log(\hat{\mathbf{y}}_i).$$

This loss provides direct supervision for labeled multimodal samples. For supervised multimodal learning, the modality fusion objective can be formulated over the training set as:

$$\begin{aligned} L_{\text{Fusion}}(\mathcal{D}^{(\text{SL})}) &= \frac{1}{n} \sum\nolimits_{(\mathbf{x}_i, \mathbf{y}_i)\in\mathcal{D}^{(\text{SL})}} \ell_{\text{Labeled}}(\mathbf{x}_i, \mathbf{y}_i), \\ &= -\frac{1}{n} \sum\nolimits_{(\mathbf{x}_i, \mathbf{y}_i)\in\mathcal{D}^{(\text{SL})}} \mathbf{y}_i^{\top} \log(\hat{\mathbf{y}}_i)), \end{aligned}$$

where, $\mathcal{D}^{(\text{SL})} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$.

Modality alignment does not rely on category labels, allowing the alignment loss $L_{\text{Align}}(\cdot)$ to be computed across the entire training set. Furthermore, as previously mentioned, we investigate prediction-level alignment,

representation-level alignment, and a hybrid alignment approach that incorporates both.

We then explore the unified integration of alignment and fusion within semi-supervised scenarios. If the category label is not available, we utilize the following loss function for optimization following classic semi-supervised learning approach [41]:

$$\ell_{\text{Unlabeled}}(\mathbf{x}_i) = -\hat{\mathbf{y}}_i^\top \log(\hat{\mathbf{y}}_i). \tag{17}$$

Eq. (17) encourages confident predictions on unlabeled samples and serves as the unsupervised component of the fusion objective.

For semi-supervised setting, labels are not available for all data. The training set is defined as: $\mathcal{D}^{(\text{SSL})} = \mathcal{D}^{(\text{L})} \bigcup \mathcal{D}^{(\text{U})}$, where $\mathcal{D}^{(\text{L})} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ and $\mathcal{D}^{(\text{U})} = \{\mathbf{x}_i\}_{i=1}^{n-m+1}$ respectively denote the labeled and unlabeled training set. Then, the objective can be formed as:

$$\begin{aligned} L_{\text{Labeled}} &= \frac{1}{|\mathcal{D}^{(\text{L})}|} \sum\nolimits_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}^{(\text{L})}} \ell_{\text{Labeled}}(\mathbf{x}_i, \mathbf{y}_i), \\ L_{\text{Unlabeled}} &= \frac{1}{|\mathcal{D}^{(\text{U})}|} \sum\nolimits_{\mathbf{x}_i\in\mathcal{D}^{(\text{U})}} \ell_{\text{Unlabeled}}(\mathbf{x}_i), \\ L_{\text{Fusion}}(\mathcal{D}^{(\text{SSL})}) &= L_{\text{Labeled}} + L_{\text{Unlabeled}}. \end{aligned} \tag{18}$$

Eq. (18) combines the supervised and unsupervised terms, allowing the fusion module to exploit both labeled and unlabeled samples.

The alignment loss remains the same as in the supervised scenario. Then, to perform model learning, we utilize an alternating algorithm between model parameters Θ and α. Specifically, our algorithmic process iteratively refines the model parameters Θ and the parameter α, employing a nested loop structure where the inner loop focuses on the currently given α to minimize total loss to update Θ, and the outer loop updates α by Eq. (13) or bi-level policy to minimize classification losses. This alternating procedure enables the model parameters and the integration weight to be optimized jointly. Through this structured optimization, the model achieves a delicate balance between fusion and alignment. The overall algorithm of our model is outlined in Algo. 1, where $\mathcal{D}$ can be substituted by $\mathcal{D}^{(\text{SL})}$ for supervised setting or $\mathcal{D}^{(\text{SSL})}$ for semi-supervised setting, and $\widetilde{\nabla}_\Theta G_{\text{Total}}(\alpha_k, \Theta_k, \zeta_t^{(1)})$ is formulated in a manner similar to $\widetilde{\nabla}_\Theta G_{\text{Fusion}}(\widehat{\Theta}_k, \zeta_k)$. Moreover, the complexity of bi-level optimization [40] and our algorithm in Algo. 1 is $\mathcal{O}(n)$, which makes our approach highly practical.

3.5 Theoretical Analysis

In this section, we present a theoretical analysis of the convergence properties of the optimization algorithm under the learning-based integration strategy. Under some assumptions, we have the following convergence result for Algo. 1. The detailed assumptions and formal theorem with its proof are included in the appendix.

Theorem 3.1 (Informal)**.** *Under some assumptions for relevant objective function and the stochastic gradient, we first fix the number of outer iterations in Algo. 1 as T and set $\eta_t^{\text{in}} = \min\left\{\frac{2}{\mu_{\text{Total}}(t+1)}, \frac{1}{\mathcal{L}_\Theta^{\text{Total}}}\right\}$, Inner_Iters $= T-1$ and $\eta_k^{\text{out}} = \frac{1}{2\mathcal{L}^*_{\text{Fusion}}\sqrt{T}}$, where $\mathcal{L}^*_{\text{Fusion}} \triangleq \frac{D_1 C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} + \frac{C_{\alpha\Theta}^{\text{Total}} \widetilde{\mathcal{L}}_{\Theta\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2} C_\Theta^{\text{Fusion}}$. Let L^*_{Fusion} is the optimal solution corresponding with the $\alpha^* \in [0,1]$ and $\bar{\alpha}_T \triangleq \frac{\sum_{t=1}^T \alpha_t}{T}$. If L_{Fusion} is convex w.r.t $\alpha \in [0,1]$, we have following convergence result for Algo. 1:*

$$\mathbb{E}[L_{\text{Fusion}}(\Theta^*(\bar{\alpha}_T))] - L^*_{\text{Fusion}} \leq \frac{1}{\sqrt{T}}\left[\mathcal{L}^*_{\text{Fusion}}|\alpha_0 - \alpha^*|^2 + 2C_1 D_1 + \frac{{\lambda_1}^2}{2\mathcal{L}^*_{\text{Fusion}}}\right],$$

where C_1, D_1 and λ_1 are all constants that only dependent on the problem parameters.

Note that η_t^{in} in the Theorem is a time-varying step size and we fix the total number of iterations of the inner loop to $T-1$. If we want to find an ϵ-stationary solution (i.e., $\exists\alpha \in [0,1]$, such that $\mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha))] - L^*_{\text{Fusion}} \leq \epsilon$) of the problem by using Algo.1, both the outer loop and the inner loop need to iterate at most $O(\frac{1}{\epsilon^2})$ times according to the theorem. Therefore, the total iteration complexity should be at most $O(\frac{1}{\epsilon^4})$. Note that the convergence rate $O(\frac{1}{T})$ of the inner loop we use to estimate $\Theta^*(\alpha)$ matches the optimal convergence rate of SGD under the strong convex smoothness assumption [42] and the overall convergence rate $O(\frac{1}{\sqrt{T}})$ of the Algo.1 matches the optimal convergence rate of SGD under the general convex smoothness assumption [43].

4 Experiments

4.1 Experimental Setup

Dataset: We select seven widely used datasets, including KSounds [44], CREMA-D [45], VGGSound [46], Sarcasm [47], Twitter2015 [48], NVGesture [27], and IEMOCAP [49] datasets, to validate our method. KSounds, CREMA-D and VGGSound datasets consist of both audio and video modalities. And Sarcasm and Twitter2015 datasets consist of both image and text modalities. NVGesture and IEMOCAP datasets are used to construct research that goes beyond the limitation of two modalities. The NVGesture dataset consists of depth, RGB, and optical flow (OF) modalities. IEMOCAP dataset consists of audio, video, and text modalities. To ensure fair comparison, all architectural choices and training hyperparameters follow the original settings of the compared methods [8, 9, 18]. The main architectural choices and training hyperparameters are summarized in Table 2. For audio–video datasets, we adopt ResNet18 backbones for both modalities. Models are trained using SGD with momentum 0.9, weight decay 10^{-1}, initial learning rate 10^{-2}, and batch size 256. The learning rate is decayed by a factor of 10 upon validation loss saturation. Video inputs consist of three uniformly sampled frames from ten-frame clips, and audio inputs are converted to spectrograms of size 257×1004 (KSounds, VGGSound) or 257×299 (CREMA-D). For image–text datasets, we employ ResNet50 for image encoding and BERT for text encoding, with images resized to 224×224 and text sequences truncated to 128 tokens. Optimization is performed using Adam with learning rate 10^{-4} and batch size 128.

We verify the effectiveness of our method under supervised and semi-supervised. In the supervised scenario, all training samples from the datasets, along with their corresponding labels, are used for model training. In the

TABLE 2
Architectural choices and training hyperparameters.

Category	Setting	Value
Architecture	Audio–Video (AV)	ResNet18
	Image–Text (IT)	ResNet50, BERT
Training	Optimizer	SGD (AV), Adam (IT)
	Initial learning rate	10^{-2} (AV), 10^{-4} (IT)
	Batch size	256 (AV), 128 (IT)
	Weight decay	10^{-1}
	Learning rate schedule	$\times 0.1$ after validation plateau
Input	Video	3 frames
	Audio	spectrogram 257×1004 or 257×299
	Image	224×224
	Text	128 tokens

semi-supervised scenario, a subset of the samples from the VGGSound training set, along with their corresponding labels, are used to construct $\mathcal{D}^{(\mathrm{SL})}$, while the remaining samples are used to construct $\mathcal{D}^{(\mathrm{SSL})}$. Implementation details are provided in the appendix.

Baselines: We select a wide range of baselines for comparison. These baselines can be divided into four groups, i.e., traditional supervised learning-based MML, rebalanced MML and semi-supervised learning-based MML. The first group encompasses techniques like feature concatenation (CONCAT), affine transformation (Affine) [50], channel-wise fusion (Channel) [33], multi-layer LSTM fusion (ML-LSTM) [51], prediction summation (Sum), prediction weighting (Weight) [52], and enhanced trust modal combination (ETMC) [53]. The second group includes MSES [20], G-Blend [8], OGM [9], DOMFN [31], MSLR [32], PMR [18], AGM [19], MLA [22], ReconBoost [24], SMV [54], DI-MML [23], and MMPareto [29]. The third group contains semi-supervised multimodal learning (SSMML) [41], including DSLR [55], SMDDRL [56], CorMulT [57], DUT [58].

Evaluation Metrics: Following [9], we use accuracy and Mean Average Precision (MAP) as evaluation metrics for audio-video datasets. For text-image datasets, we adopt accuracy and Macro F1-score (MacroF1) [47]. Accuracy measures the proportion of correct predictions to total predictions, indicating the overall predictive accuracy. Macro F1 calculates the average of F1 scores across all categories, balancing precision and recall to evaluate performance evenly across classes. MAP represents the average precision across all categories.

4.2 Main Experimental Results

Comparison under Supervised Learning Scenario: We first compare our method under supervised learning setting. The results on audio-video datasets, i.e., KSounds, CREMA-D and VGGSound datasets, and image-text datasets, i.e., Sarcasm and Twitter2015 datasets, are reported in Table 3. In Table 3, the best and second best performance are denoted with bold and underlining, respectively. Furthermore, we utilize symbol † to denote the results which are based on MML but perform worse than the best unimodal approach.

According to the results in Table 3, we can draw the following observations: (1). Compared with all baselines including traditional multimodal learning approaches and fusion methods with modal rebalancing strategies, our method with learning-based strategy can achieve best performance by a large margin in almost all cases. We can also find that the model with learning-based strategy can achieve better performance than that with heuristic strategy. (2). Across the Twitter2015 dataset, there is a discernible trend where the optimal unimodal performance outstrips that of multimodal joint learning. Additionally, in other datasets, fusion methodologies devoid of rebalancing mechanisms manifest negligible enhancements relative to the foremost unimodal performance, notably on the CREMA-D and Sarcasm datasets. This shortfall originates from the prevalent challenge of modality imbalance. (3). Every modality rebalancing technique demonstrates significant improvements over traditional feature concatenation fusion. This finding not only underscores the detrimental impact of modality imbalance on performance but also corroborates the efficacy of the modality rebalancing approach.

Furthermore, we report the results on NVGesture and IEMOCAP dataset which contain three modalities in Table 4, where some baselines including DOMEN [31] and OGM [9] are ignored since they cannot be applied to the case with three modalities, and the unimodal results are based on RGB/OF/Depth and audio/video/text for NVGesture and IEMOCAP datasets, respectively. We can see that differing from modal rebalancing methods restricted to scenarios with only two modalities, our approach with learning-based strategy can address challenges in scenarios involving more than two modalities and achieve the best results on both NVGesture and IEMOCAP datasets. Furthermore, our method with heuristic strategy can also outperform all baselines in almost all cases.

Comparison under Semi-Supervised Learning Scenario: We further provide the results under semi-supervised setting by comparing our method with three groups of baselines, i.e., traditional supervised learning-based MML baselines, rebalanced multimodal baselines, and semi-supervised multimodal learning baselines. Traditional supervised learning-based MML baselines include Concat and Sum approaches. For rebalanced multimodal learning, we select OGM [9], MMPareto [29], SMV [54], MLA [22], ReconBoost [24], DI-MML [23] for comparison. Semi-supervised MML baselines include DSLR [55], SMDDRL [56], CorMulT [57], DUT [58]. We randomly select a subset of samples with a proportion of μ to construct the unlabeled training set, where μ is set to be 40%. For traditional supervised multimodal learning baselines and rebalanced baselines, in addition to the original algorithm, we modify their classification loss function in a semi-supervised manner like Eq. (17) to handle the unlabeled training data. The corresponding approaches are denoted with the suffix "+SSL". Furthermore, we report the results with different μ values in appendix.

From the results in Table 5, we can draw the following observations: (1). Our method can achieve the best performance in most cases by comparing with the baselines including rebalanced MML approaches and semi-supervised MML approach. (2). Compared with vanilla MML, rebalanced MML baselines and semi-supervised baselines can achieve better performance in all cases, demonstrating

TABLE 3
Comparison with SOTA multimodal learning methods. The best results are highlighted in bold. The underlining symbol denotes the second best performance. We use the symbol † to denote the results which are based on MML but perform worse than the best unimodal approach. We also report the performance gain (or degradation) of learning-based method compared to the best-performing baseline. The proposed methods consistently achieve competitive or superior performance across datasets.

Method	KSounds		CREMA-D		VGGSound		Sarcasm		Twitter2015	
	Accuracy	MAP	Accuracy	MAP	Accuracy	MAP	Accuracy	MacroF1	Accuracy	MacroF1
Audio/Image	54.12%	56.69%	45.83%	58.79%	46.55%	47.01%	71.81%	70.73%	58.63%	43.33%
Video/Text	55.62%	58.37%	63.17%	68.61%	34.94%	34.78%	81.36%	80.56%	73.67%	68.49%
Concat	64.55%	71.30%	63.61%	68.41%†	51.16%	53.52%	82.86%	82.40%	70.11%†	63.86%†
Affine	64.24%	69.31%	66.26%	71.93%	50.01%	51.55%	82.40%	81.88%	72.03%†	59.92%†
Channel	64.51%	68.66%	66.13%	71.70%	51.10%	53.05%	-	-	-	-
ML-LSTM	63.94%	69.02%	62.90%†	64.73%†	49.66%	51.39%	82.77%	82.05%	70.68%†	65.64%†
Sum	64.90%	71.03%	63.44%	69.08%	51.36%	53.38%	82.94%	82.47%	73.00%†	66.61%†
Weight	65.33%	71.10%	66.53%	71.34%	51.44%	53.00%	82.65%	82.19%	72.42%†	65.16%†
ETMC	65.67%	72.50%	65.86%	71.34%	51.15%	53.38%	83.69%	83.23%	73.96%	67.39%†
MSES	65.91%	71.96%	65.46%	71.38%	48.91%	54.29%	84.23%	83.69%	71.84%†	66.55%†
G-Blend	67.22%	72.74%	64.65%	73.92%	50.96%	55.55%	82.86%	82.15%	74.35%	68.69%
MSLR	67.56%	72.82%	68.68%	74.12%	49.87%	54.15%	84.39%	83.78%	72.52%†	64.39%†
DOMFN	66.25%	72.44%	67.34%	73.72%	49.43%	53.77%	83.56%	82.62%	74.45%	68.57%
OGM	65.82%	71.59%	66.12%	73.72%	48.29%	49.78%	83.60%	82.93%	**74.92%**	68.74%
PMR	66.75%	72.74%	66.59%	70.58%	46.47%	48.66%	83.10%	82.56%	74.25%	68.62%
AGM	67.91%	73.88%	67.33%	78.07%	47.11%	51.98%	83.06%	82.93%	74.83%	69.11%
MMPareto	70.00%	78.50%	74.87%	85.35%	51.25%	54.73%	83.48%	82.84%	73.58%†	67.29%†
SMV	69.00%	74.26%	78.72%	84.17%	50.31%	53.62%	84.18%	83.68%	74.28%	68.17%
MLA	70.04%	79.45%	79.43%	85.72%	51.65%	54.73%	84.26%	83.48%	73.52%†	67.13%†
ReconBoost	68.55%	76.62%	75.57%	81.40%	50.97%	53.87%	84.37%	83.17%	74.42%	68.34%
DI-MML	72.03%	74.26%	81.58%	85.92%	51.73%	54.79%	84.11%	83.15%	72.48%†	66.86%†
Ours-H	71.78%	78.83%	82.93%	89.64%	51.63%	54.08%	84.39%	83.77%	73.77%	68.52%
Ours-LB	**73.33%**	**80.24%**	**84.27%**	**91.84%**	**54.46%**	**58.98%**	**84.99%**	**84.45%**	74.86%	**69.78%**
Δ	↑1.30%	↑0.79%	↑2.69%	↑5.92%	↑2.73%	↑3.43%	↑0.60%	↑0.77%	↓0.06%	↑0.67%

TABLE 4
Performance comparison on the datasets with three modalities. The results show that the proposed method generalizes effectively to the three-modality setting and achieves superior overall performance.

Method	NVGesture		IEMOCAP	
	Accuracy	MacroF1	Accuracy	MAP
RGB/Audio	78.22%	78.33%	58.45%	62.75%
OF/Video	78.63%	78.65%	30.71%	29.41%
Depth/Text	81.54%	81.83%	70.55%	78.30%
MSES	81.12%	81.47%	71.73%	79.90%
G-Blend	82.99%	83.05%	70.10%	77.38%
MSLR	82.86%	82.92%	76.69%	85.06%
AGM	82.78%	82.82%	77.51%	84.71%
MMPareto	83.82%	84.24%	77.69%	85.91%
SMV	83.52%	83.41%	78.32%	85.48%
MLA	83.73%	83.87%	79.31%	**86.83**%
ReconBoost	84.13%	**86.32**%	76.87%	84.49%
Ours-H	85.06%	85.15%	78.97%	85.43%
Ours-LB	**85.48**%	85.65%	**79.96**%	86.04%
Δ	↑1.35%	↓0.67%	↑0.65%	↓0.79%

the effectiveness of rebalanced multimodal learning and semi-supervised learning, respectively. (3). Compared with semi-supervised baselines, rebalanced MML baselines can achieve competitive performance, demonstrating the effectiveness of rebalanced multimodal learning.

4.3 Ablation Study

Impact of Modality Alignment and Dynamic Integration: To comprehensively assess the effectiveness of our method, we conduct experiments to study the influence of main components, i.e., modality alignment and dynamic integration (DI). The results are shown in Table 6, where the "PA", "RA" and "DI" denote that whether the prediction-level, representation-level alignment and dynamic integration are applied during training. The unimodal MAP results are based on audio and video modalities for KSounds and CREMA-D datasets, and unimodal F1 results are based on image and text modalities for Sarcasm , VGGSound and Twitter2015 datasets. Please note that dynamic integration depends on the modality alignment. Hence the method with dynamic integration but without modality alignment cannot be performed. From Table 6, we can see that both modality alignment and dynamic integration can boost performance in multimodal learning.

Analysis of Integration Strategy: Integration weight α balances modality fusion and alignment. A fixed α is suboptimal, since alignment is more critical in early training to reduce cross-modal discrepancies, while fusion becomes more important in later stages for improving task performance. Dynamically adjusting α is therefore necessary. Specifically, we analyze three categories of integration strategy, i.e., constant, stepwise and dynamic strategies. For constant strategy, we assign a constant value to α and run three sets of experiments by setting $\alpha = 0$, $\alpha = 0.5$, and $\alpha = 1$. Here, $\alpha = 0$ denotes that we only perform modality fusion. $\alpha = 1$ is defined similarly. For stepwise strategy, we define an indicator function $h(p)$, where $h(p)$ denotes that $\alpha = p$ if the current epoch is less than half of the total epochs, otherwise $\alpha = 1 - p$. For dynamic strategy, we assign value to α by heuristic strategy (Ours-H) and learning-based strategy (Ours-LB).

The experimental results in Table 7 show that: (1). Integrating modality fusion and modality alignment simultaneously with a constant ratio can slightly boost performance in some cases and greatly reduce the performance gap between audio and video. (2). In general, the model with a

TABLE 5
Comparison with representative multimodal learning methods under the semi-supervised setting ($\mu = 40\%$). The best and the second best results are indicated as bold and underlying. We also report the performance gain (or degradation) of learning-based method compared to the best-performing baseline. The proposed method achieves the best or highly competitive performance across the five benchmark datasets.

Method	KSounds		CREMA-D		VGGSound		Sarcasm		Twitter2015	
	Accuracy	MAP	Accuracy	MAP	Accuracy	MAP	Accuracy	MacroF1	Accuracy	MacroF1
Concat	53.77%	62.59%	54.30%	69.38%	34.43%	36.58%	80.41%	80.10%	69.62%	64.19%
Sum	52.69%	61.03%	53.09%	64.56%	33.62%	36.14%	80.12%	79.53%	68.08%	59.98%
Concat+SSL	54.93%	63.10%	55.11%	70.27%	35.04%	36.73%	81.57%	80.41%	69.91%	61.90%
Sum+SSL	54.12%	62.14%	54.97%	68.91%	34.08%	36.35%	81.36%	81.02%	68.47%	53.73%
OGM	56.20%	63.85%	56.85%	70.93%	34.91%	36.35%	80.95%	80.35%	72.89%	64.38%
MMPareto	58.37%	67.09%	64.65%	80.82%	35.14%	36.91%	81.40%	80.66%	71.94%	65.52%
SMV	59.03%	70.35%	68.28%	83.78%	34.98%	36.79%	81.24%	80.67%	72.23%	65.36%
MLA	62.31%	71.18%	69.35%	78.10%	36.11%	37.39%	81.78%	81.15%	71.75%	65.93%
ReconBoost	61.58%	71.40%	65.32%	82.92%	36.78%	37.55%	81.57%	81.18%	72.03%	65.06%
DI-MML	61.65%	71.07%	71.51%	83.56%	36.30%	38.91%	81.74%	81.28%	71.65%	64.09%
OGM+SSL	57.09%	64.71%	57.66%	69.02%	34.79%	37.33%	81.36%	81.02%	73.07%	66.31%
MMPareto+SSL	60.49%	70.60%	65.46%	81.80%	36.11%	37.39%	82.15%	81.71%	72.13%	66.20%
SMV+SSL	61.19%	69.42%	69.62%	82.99%	35.40%	37.86%	82.03%	81.40%	72.42%	66.50%
MLA+SSL	63.36%	72.54%	71.77%	78.75%	39.14%	40.77%	82.36%	81.84%	72.23%	64.53%
ReconBoost+SSL	62.47%	69.42%	67.47%	83.19%	37.46%	39.79%	82.40%	81.89%	72.32%	64.77%
DI-MML+SSL	62.43%	72.84%	72.98%	84.46%	37.43%	40.43%	82.27%	81.86%	72.42%	64.35%
DSLR	62.08%	71.44%	72.58%	81.52%	39.08%	40.78%	82.15%	81.12%	71.07%	63.65%
SMDDRL	61.58%	71.40%	73.12%	83.51%	39.43%	**41.04%**	82.52%	81.93%	72.61%	65.20%
CorMulT	63.63%	73.32%	74.06%	80.90%	38.53%	40.50%	81.94%	81.53%	71.55%	64.35%
DUT	64.01%	**73.38%**	74.19%	**85.95%**	38.92%	40.72%	82.86%	82.10%	72.81%	65.54%
Ours-LB	63.36%	72.54%	73.66%	82.02%	37.82%	40.97%	81.86%	81.19%	71.84%	64.55%
Ours-LB+SSL	**64.67%**	73.04%	**74.33%**	84.41%	**39.53%**	41.02%	**82.94%**	**82.31%**	**73.19%**	**65.83%**
Δ	↑0.66%	↓0.34%	↑0.14%	↓1.54%	↑0.10%	↓0.02%	↑0.08%	↑0.21%	↑0.38%	↑0.29%

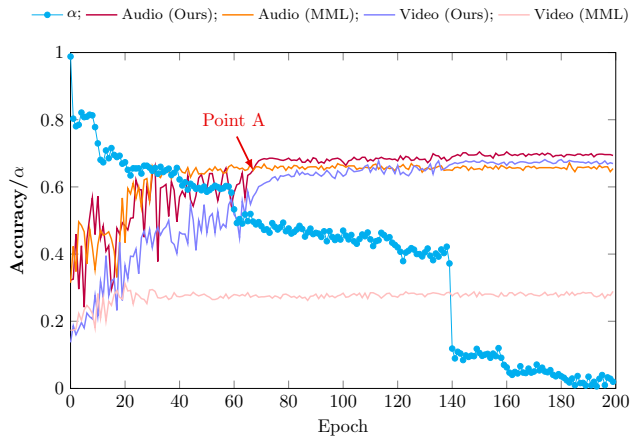


Fig. 3. Evolution of α on the CREMA-D dataset under the learning-based integration strategy. The learned α gradually shifts from alignment-oriented training to fusion-oriented training, indicating that effective optimization first emphasizes cross-modal alignment and then increases the contribution of task-oriented fusion.

stepwise strategy can outperform the model with a constant strategy. Furthermore, we can find that the performance of the model with two-stage training, i.e., $h(0)$ or $h(1)$, is worse than that of the model which combines two losses with a constant value, i.e., $h(0.05)$ or $h(0.95)$. (3). The overall performance of the model with the dynamic strategy is better than that with the other strategy. Moreover, the model with the learning-based strategy can achieve the best performance. And the performance gap of this model is nil or negligible. These results indicate that reducing the unimodal performance gap is closely related to improving multimodal performance.

Change of α for Learning-based Strategy: To further observe the change of the optimal α during training, we illustrate the change of α on CREMA-D dataset. The results are shown in Figure 3, where α is calculated by learning-based strategy. From Figure 3, we can see that at the initial stage of training, the model focuses on aligning the modalities. As training progresses, the importance of modal fusion increases, leading to improved overall performance.

Positive Impact Analysis of Alignment: Figure 3 further illustrates the unimodal performance of our method, i.e., Audio (Ours) and Video (Ours), compared to that of general multimodal learning approach, i.e., Audio (MML) and Video (MML). Building on the experimental results in Figure 3, we analyze the positive impact of alignment on the dominant modality: Due to the imposed modality alignment, the performance of the dominant modality is not as good as the vanilla MML at the beginning of training (about before the 65-th epoch, which we mark with "*Point A*" in Figure 3). As the training progresses, α further decreases, and the performance of the dominant modality also improves further, and it clearly exceeds the effect of the vanilla MML training method.

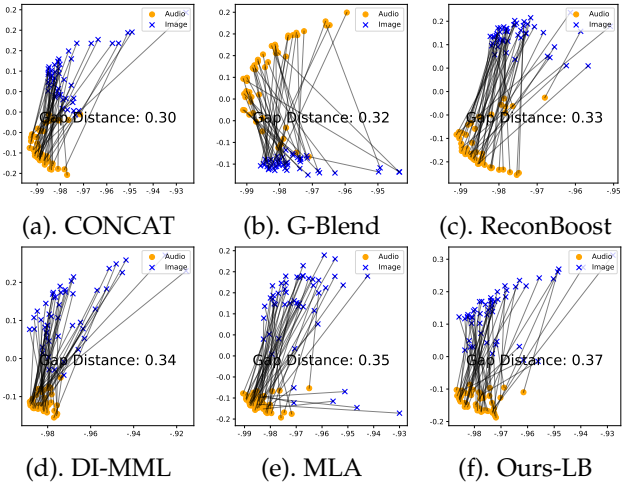
4.4 Further Analysis

Analysis of Modality Gap: Modality gap [59] characterizes the correlation between different modalities in multimodal learning. And large modality gap leads to better performance in some situations. We further illustrate the modality gap for CONCAT, G-Blend, MLA, and our method. The results in Figure 4 show that our method can learn more discriminative representations and results in higher accuracy with a large modality gap compared with other methods.

Robustness Analysis of the Pretrained Model: We further exploit the robustness of the large-scale language vision

TABLE 6
Ablation study. The symbols "PA", "RA" and "DI" denote whether the prediction alignment, representation alignment, and dynamic integration are applied during training. Results show that each component improves multimodal learning performance and their combination achieves the best results.

Dataset	Module			MAP/F1		
	PA	RA	DI	Multiple	Unimodal1	Unimodal2
KSounds	×	×	×	69.32%	48.82%	27.19%
	✓	×	×	74.46%	54.34%	57.22%
	✓	×	✓	78.83%	57.82%	60.15%
	×	✓	×	71.76%	51.05%	47.05%
	×	✓	✓	78.97%	58.40%	60.42%
	✓	✓	×	76.69%	57.10%	59.77%
	✓	✓	✓	80.24%	58.20%	60.50%
CREMA-D	×	×	×	76.07%	70.97%	34.15%
	✓	×	×	89.19%	73.62%	73.44%
	✓	×	✓	90.00%	75.72%	73.81%
	×	✓	×	86.32%	72.11%	52.51%
	×	✓	✓	90.06%	75.27%	67.36%
	✓	✓	×	89.32%	75.79%	74.58%
	✓	✓	✓	91.84%	76.27%	76.12%
VGGSound	×	×	×	48.84%	38.99%	7.73%
	✓	×	×	56.26%	46.96%	34.98%
	✓	×	✓	56.61%	44.44%	28.48%
	×	✓	×	56.09%	47.16%	34.58%
	×	✓	✓	55.98%	45.93%	34.36%
	✓	✓	×	56.52%	47.21%	34.83%
	✓	✓	✓	58.95%	48.04%	34.75%
Sarcasm	×	×	×	82.43%	62.81%	77.96%
	✓	×	×	82.85%	71.06%	81.08%
	✓	×	✓	83.62%	71.99%	82.44%
	×	✓	×	83.10%	68.74%	80.72%
	×	✓	✓	84.57%	74.53%	83.03%
	✓	✓	×	83.18%	73.02%	82.04%
	✓	✓	✓	84.45%	73.18%	82.14%
Twitter2015	×	×	×	63.86%	40.99%	68.38%
	✓	×	×	64.89%	48.47%	67.53%
	✓	×	✓	69.17%	51.33%	68.88%
	×	✓	×	65.33%	46.84%	69.04%
	×	✓	✓	70.57%	53.43%	69.68%
	✓	✓	×	67.78%	51.68%	68.67%
	✓	✓	✓	70.59%	52.03%	69.78%



(a). CONCAT (b). G-Blend (c). ReconBoost

(d). DI-MML (e). MLA (f). Ours-LB

Fig. 4. Visualizations of the modality gap on the CREMA-D dataset. Compared with existing methods, Ours-LB exhibits a larger modality gap together with higher accuracy, suggesting that the proposed method learns more discriminative multimodal representations.

pretrained model CLIP [38] model on Sarcasm and Twitter2015 datasets. We replace the encoders for image and text as the corresponding encoders pretrained by CLIP and fine-tune the model on Sarcasm and Twitter2015 datasets respectively. The results are shown in Table 8, where "CLIP+MLA" and "CLIP+Ours-LB" present that we apply the MLA's

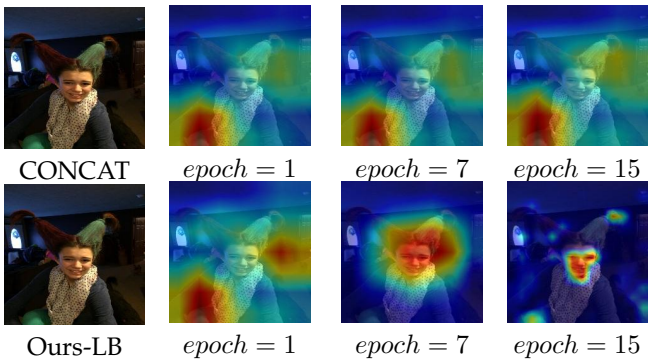


Fig. 5. Grad-CAM visualizations on the Twitter2015 dataset for CONCAT and Ours-LB at different training epochs. Compared with CONCAT, Ours-LB first promotes cross-modal alignment and then focuses on task-relevant regions, illustrating the progressive transition from alignment-driven optimization to classification-oriented learning.

and our algorithm, respectively. From Table 8, we can draw the following observations: (1). Both CLIP+MLA and CLIP+Ours-LB can outperform CLIP in all cases. (2). With the help of dynamic integration, the performance of our method is better than that of MLA.

Visualization: We utilize GradCAM [60] to showcase the visualization of image regions that attract the weak modality's focus during training. By using GradCAM, the importance scores are assigned to every pixel in each feature map, aiding in identifying the image regions critical for the model's predictions. We compare the visualization for CONCAT and our method. The visualization results are presented in Figure 5, where the second, the third, and the last columns denote the results of the first, the seventh, and the last epoch, respectively. The category label for this image is "Negative" and the corresponding text is "Crazy hair day ! T is a contender.". By comparing our method with CONCAT, we can see that our method focuses on the textual information from text modality, and then fits the learned features to the category labels.

5 Conclusion

In this paper, we investigate modality imbalance in multimodal learning from a unified perspective of modality fusion and dynamic modality alignment. We observe that task-oriented fusion tends to favor dominant modalities when cross-modal predictions or representations are insufficiently aligned, which limits the exploitation of weak modalities and leads to imbalanced multimodal learning. To address this issue, we propose a unified framework that dynamically incorporates modality alignment into task-oriented fusion at the prediction, representation, and hybrid levels, thereby adaptively enhancing cross-modal consistency and stimulating weak modality capabilities during multimodal learning. Since the desired strength of alignment varies across training stages, with stronger alignment being necessary to reduce cross-modal discrepancy in the early stage and excessive alignment potentially constraining discriminative fusion in the later stage, we further develop heuristic and learning-based strategies to adaptively control the alignment effect within the fusion process. The convergence of the learning-based optimization algorithm

TABLE 7
Comparison of dynamic integration strategy on KSounds and CREMA-D datasets. Dynamic integration achieves the best overall performance while yielding a negligible unimodal performance gap, indicating that adaptively balancing fusion and alignment is more effective than fixed or manually scheduled strategies.

Dataset	Modality	Constant			Stepwise				Dynamic	
		0	0.5	1	$h(0)$	$h(1)$	$h(0.05)$	$h(0.95)$	Ours-H	Ours-LB
KSounds	Multiple	64.67%	70.51%	28.37%	68.11%	69.42%	68.96%	70.76%	71.78%	73.75%
	Audio	48.78%	52.65%	29.65%	49.67%	52.49%	52.53%	52.65%	54.35%	55.35%
	Video	33.01%	56.67%	21.88%	53.69%	54.27%	56.40%	55.51%	56.40%	56.78%
CREMA-D	Multiple	70.97%	82.39%	28.90%	81.18%	82.53%	81.85%	82.66%	82.93%	84.27%
	Audio	65.46%	69.62%	26.75%	68.41%	68.82%	67.88%	68.15%	70.56%	69.62%
	Video	28.90%	64.92%	23.25%	61.02%	66.40%	63.04%	66.53%	65.73%	67.88%

TABLE 8
Results on the Sarcasm and Twitter2015 datasets achieved by using the CLIP pretrained model as encoders. CLIP+Ours-LB achieves the best overall performance, demonstrating the compatibility of the proposed method with pretrained multimodal encoders.

Dataset	Method	Multiple	Image	Text
Sarcasm	CLIP	83.11%	74.82%	82.15%
	CLIP+MLA	84.45%	77.45%	83.19%
	CLIP+Ours-LB	85.64%	78.46%	83.77%
Twitter2015	CLIP	72.52%	54.48%	71.75%
	CLIP+MLA	73.95%	56.53%	72.37%
	CLIP+Ours-LB	74.87%	57.70%	73.67%

is theoretically proved to ensure its reliability. Extensive experiments under supervised and semi-supervised scenarios demonstrate that our method consistently outperforms state-of-the-art multimodal learning approaches, achieving accuracy improvements of 1.30%, 2.69%, and 0.60% on representative bimodal benchmarks, namely KSounds, CREMA-D, and Sarcasm, respectively, as well as gains of 1.35% and 0.65% on trimodal datasets, namely NVGesture and IEMOCAP.

While the proposed framework demonstrates strong empirical performance, several limitations remain and open directions for future research. Here, we discuss several limitations of the current approach and outline potential directions for future research.

(1). Reliance on Well-Structured Modality Pairs: The current framework assumes that multimodal inputs are reasonably structured and that cross-modal relationships are stable. In real-world scenarios, modalities may be weakly correlated, partially missing, or exhibit temporal misalignment, which can undermine the effectiveness of the proposed alignment mechanisms. *Future Work:* Develop adaptive alignment strategies that can handle modality uncertainty, missing modalities, and asynchronous data streams, possibly through uncertainty estimation or dynamic modality weighting.

(2). Scalability to More Than Two Modalities: Although the method is conceptually extensible to multiple modalities, the interaction complexity between fusion and alignment grows significantly as the number of modalities increases. The current design does not explicitly address scalability or computational efficiency in such settings. *Future Work:* Investigate scalable fusion-alignment architectures, such as modular or hierarchical alignment mechanisms, and explore efficient optimization strategies tailored for high-dimensional multimodal spaces.

(3). Limited Theoretical Analysis on Alignment Regularization: The paper provides convergence guarantees for the learning-based integration strategy but does not theoretically analyze how alignment regularization affects the generalization error or the optimization dynamics of dominant versus weak modalities. *Future Work:* Establish theoretical bounds on generalization error under alignment constraints and analyze the optimization landscape to better understand the trade-offs between fusion and alignment.

ACKNOWLEDGMENT

This work was supported in part by the NSFC (62276131, 62506168), in part by the Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081, BK20251431), and in part by the Fundamental Research Funds for the Central Universities (No.30925010205).

REFERENCES

[1] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.

[2] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *TPAMI*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[3] A. X. Chang, W. Monroe, M. Savva, C. Potts, and C. D. Manning, "Text to 3d scene generation with rich lexical grounding," in *ACL*, 2015, pp. 53–62.

[4] L. Zhu, T. Wang, F. Li, J. Li, Z. Zhang, and H. T. Shen, "Cross-modal retrieval: A systematic review of methods and future directions," *CoRR*, vol. abs/2308.14263, 2023.

[5] Y. Yang, J. Guo, G. Li, L. Li, W. Li, and J. Yang, "Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning," *FCS*, vol. 18, no. 3, p. 181335, 2024.

[6] L. Chen, B. Li, S. Shen, J. Yang, C. Li, K. Keutzer, T. Darrell, and Z. Liu, "Large language models are visual reasoning coordinators," in *NeurIPS*, 2023, pp. 70 115–70 140.

[7] W. Liu, I. W. Tsang, and K. Müller, "An easy-to-hard learning paradigm for multiple classes and multiple labels," *JMLR*, vol. 18, pp. 1–38, 2017.

[8] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*, 2020, pp. 12 692–12 702.

[9] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8228–8237.

[10] Y. Yang, H. Ye, D. Zhan, and Y. Jiang, "Auxiliary information regularized machine for multiple modality feature learning," in *IJCAI*, 2015, pp. 1033–1039.

[11] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," in *NeurIPS*, 2021, pp. 10 944–10 956.

[12] 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 *ICML*, 2022, pp. 24 043–24 055.

[13] 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 *ICML*, 2022, pp. 9226–9259.

[14] J. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," *CoRR*, vol. abs/2006.15955, 2020.

[15] J. Yang, C. Dai, D. Ou, D. Li, J. Huang, D. Zhan, X. Zeng, and Y. Yang, "COURIER: contrastive user intention reconstruction for large-scale visual recommendation," *FCS*, vol. 19, no. 7, p. 197602, 2025.

[16] H. Zhang, Y. Li, and X. Li, "Constrained bipartite graph learning for imbalanced multi-modal retrieval," *TMM*, vol. 26, pp. 4502–4514, 2024.

[17] C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, and H. Zhao, "On uni-modal feature learning in supervised multi-modal learning," in *ICML*, 2023, pp. 8632–8656.

[18] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "PMR: prototypical modal rebalance for multimodal learning," in *CVPR*, 2023, pp. 20 029–20 038.

[19] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *ICCV*, 2023, pp. 22 157–22 167.

[20] N. Fujimori, R. Endo, Y. Kawai, and T. Mochizuki, "Modality-specific learning rate control for multimodal classification," in *ACPR*, vol. 12047, 2019, pp. 412–422.

[21] Y. Yang, H. Pan, Q. Jiang, Y. Xu, and J. Tang, "Learning to rebalance multi-modal optimization by adaptively masking subnetworks," *TPAMI*, vol. 47, no. 6, pp. 4553–4566, 2025.

[22] X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," in *CVPR*, 2024, pp. 27 446–27 456.

[23] Y. Fan, W. Xu, H. Wang, J. Liu, and S. Guo, "Detached and interactive multimodal learning," in *ACMMM*, 2024, pp. 5470–5478.

[24] C. Hua, Q. Xu, S. Bao, Z. Yang, and Q. Huang, "Reconboost: Boosting can achieve modality reconcilement," in *ICML*, 2024, pp. 19 573–19 597.

[25] Q. Jiang, Z. Chi, and Y. Yang, "Interactive multimodal learning via flat gradient modification," in *IJCAI*, 2025, pp. 5489–5497.

[26] Q. Jiang, L. Huang, and Y. Yang, "Rethinking multimodal learning from the perspective of mitigating classification ability disproportion," in *NeurIPS*, 2025, pp. 129 580–129 607.

[27] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *CVPR*, 2016, pp. 4207–4215.

[28] Y. Yang, X. Wu, and Q. Jiang, "Towards equilibrium: An instantaneous probe-and-rebalance multimodal learning approach," in *IJCAI*, 2025, pp. 3552–3560.

[29] Y. Wei and D. Hu, "Mmpareto: Boosting multimodal learning with innocent unimodal assistance," in *ICML*, 2024, pp. 52 559–52 572.

[30] Y. Yang, F. Wan, Q.-Y. Jiang, and Y. Xu, "Facilitating multimodal classification via dynamically learning modality gap," in *NeurIPS*, 2024, pp. 1–15.

[31] Y. Yang, J. Zhang, F. Gao, X. Gao, and H. Zhu, "DOMFN: A divergence-orientated multi-modal fusion network for resume assessment," in *ACMMM*, 2022, pp. 1612–1620.

[32] Y. Yao and R. Mihalcea, "Modality-specific learning rates for effective multimodal additive late-fusion," in *ACL*, 2022, pp. 1824–1834.

[33] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: multimodal transfer module for CNN fusion," in *CVPR*, 2020, pp. 13 286–13 296.

[34] Y. Li, H. Ding, Y. Lin, X. Feng, and L. Chang, "Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis," *AIR*, vol. 57, no. 4, p. 78, 2024.

[35] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1935.

[36] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[37] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019, pp. 13–23.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[39] L. N. Vicente and P. H. Calamai, "Bilevel and multilevel programming: A bibliography review," *JGO*, vol. 5, no. 3, pp. 291–306, 1994.

[40] S. Ghadimi and M. Wang, "Approximation methods for bilevel programming," *CoRR*, vol. abs/1802.02246, 2018.

[41] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *NeurIPS*, 2004, pp. 529–536.

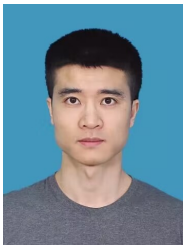
[42] E. Hazan and S. Kale, "Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization," in *COLT*, vol. 19, 2011, pp. 421–436.

[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAMJO*, vol. 19, no. 4, pp. 1574–1609, 2009.

[44] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017, pp. 609–617.

[45] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: crowd-sourced emotional multimodal actors dataset," *TAC*,

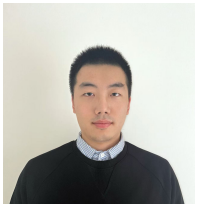
vol. 5, no. 4, pp. 377–390, 2014.
[46] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020, pp. 721–725.
[47] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *ACL*, 2019, pp. 2506–2515.
[48] J. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *IJCAI*, 2019, pp. 5408–5414.
[49] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *LRE*, vol. 42, no. 4, pp. 335–359, 2008.
[50] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018, pp. 3942–3951.
[51] W. Nie, Y. Yan, D. Song, and K. Wang, "Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition," *MTA*, vol. 80, no. 11, pp. 16 205–16 214, 2021.
[52] Y. Yang, K. Wang, D. Zhan, H. Xiong, and Y. Jiang, "Comprehensive semi-supervised multi-modal learning," in *IJCAI*, 2019, pp. 4092–4098.
[53] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *TPAMI*, vol. 45, no. 2, pp. 2551–2566, 2023.
[54] Y. Wei, R. Feng, Z. Wang, and D. Hu, "Enhancing multimodal cooperation via sample-level modality valuation," in *CVPR*, 2024, pp. 27 338–27 347.
[55] S. Wang, Z. Chen, S. Du, and Z. Lin, "Learning deep sparse regularizers with applications to multi-view clustering and semi-supervised classification," *TPAMI*, vol. 44, no. 9, pp. 5042–5055, 2022.
[56] X. Jia, X. Jing, X. Zhu, S. Chen, B. Du, Z. Cai, Z. He, and D. Yue, "Semi-supervised multi-view deep discriminant representation learning," *TPAMI*, vol. 43, no. 7, pp. 2496–2509, 2021.
[57] Y. Li, R. Zhu, and W. Li, "Cormult: A semi-supervised modality correlation-aware multimodal transformer for sentiment analysis," *CoRR*, vol. abs/2407.07046, 2024.
[58] X. Wan, J. Liu, X. Liu, Y. Wen, H. Yu, S. Wang, S. Yu, T. Wan, J. Wang, and E. Zhu, "Decouple then classify: A dynamic multi-view labeling strategy with shared and specific information," in *ICML*, 2024, pp. 49 941–49 956.
[59] 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," in *NeurIPS*, 2022, pp. 17 612–17 625.
[60] 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," *IJCV*, vol. 128, no. 2, pp. 336–359, 2020.



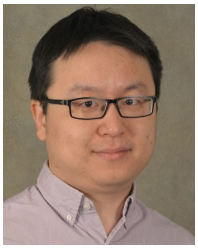
Yang Yang received the Ph.D. degree in computer science, Nanjing University, China in 2019. At the same year, he became a faculty member at Nanjing University of Science and Technology, China. He is currently a Professor with the school of Computer Science and Engineering. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining. He has published prolifically in refereed journals and conference proceedings, including IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Information Systems (ACM TOIS), ACM Transactions on Knowledge Discovery from Data (TKDD), ACM SIGKDD, ACM SIGIR, WWW, IJCAI, and AAAI. He was the recipient of the Best Paper Award of ACML-2017. He serves as PC in leading conferences such as IJCAI, AAAI, ICML, NeurIPS, etc.

Fengqiang Wan is currently working towards the Ph.D. degree at the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly lie in deep learning and data mining. He is currently focusing on multi-modal learning.

Qing-Yuan Jiang received the BSc and the PhD degrees in computer science from Nanjing University, China. He has published more than ten papers in leading international journals/conferences. He serves as a PC/Reviewer in leading conferences such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Image Processing (TIP), NeurIPS, ICML, ICLR, AISTATS, AAAI, IJCAI, etc. He is currently an associate professor at Nanjing University of Science and Technology. His research interests are in machine learning, multi-modal learning, and information retrieval.



Tianxi Zhu is currently a graduate student at the School of Computer Science and Technology, Dalian University of Technology. His research interests focus on the application of large-scale stochastic optimization in machine learning.



Yi Xu received the PhD degree in computer science from the University of Iowa, Iowa City, Iowa, USA. He is currently a professor with the School of Control Science and Engineering, Dalian University of Technology, China. His research interests are machine learning, optimization, deep learning, and statistical learning theory. He has published more than twenty papers in refereed journals and conference proceedings, including IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Transactions on Machine Learning Research, NeurIPS, ICML, ICLR, CVPR, AAAI, IJCAI, UAI. He serves as a SPC/PC/Reviewer in leading conferences such as NeurIPS, ICML, ICLR, CVPR, AAAI, IJCAI, etc.