Intra- and Inter-Modal Curriculum for Multimodal Learning

Yuwei Zhou
Department of Computer Science and Technology, Tsinghua University
zhou-yw21@mails.tsinghua.edu.cn

Xin Wang*
Department of Computer Science and Technology, BNRist, Tsinghua University
xin_wang@tsinghua.edu.cn

Hong Chen
Department of Computer Science and Technology, Tsinghua University
h-chen20@mails.tsinghua.edu.cn

Xuguang Duan
Department of Computer Science and Technology, Tsinghua University
dxg18@mails.tsinghua.edu.cn

Wenwu Zhu*
Department of Computer Science and Technology, BNRist, Tsinghua University
wwzhu@tsinghua.edu.cn

ABSTRACT

Multimodal learning has been widely studied and applied due to its improvement over previous unimodal tasks and its effectiveness on emerging multimodal challenges. However, it has been reported that modal encoders are under-optimized in multimodal learning in contrast to unimodal learning, especially when some modalities are dominant over others. Existing solutions to this problem suffer from two limitations: i) they merely focus on inter-modal balance, failing to consider the influence of intra-modal data on each modality; ii) their implementations heavily rely on unimodal performances or losses, thus being suboptimal for the tasks requiring modal interactions (e.g., visual question answering). To tackle these limitations, we propose I^{2}MCL, a generic **I**ntra- and **I**nter-**M**odal **C**urriculum **L**earning framework which simultaneously considers both data difficulty and modality balance for multimodal learning. In the intra-modal curriculum, we adopt a pretrained teacher model to obtain knowledge distillation loss as the difficulty measurer, which determines the data weights within the corresponding modality. In the inter-modal curriculum, we utilize a Pareto optimization strategy to measure and compare the gradients from distillation loss and task loss across modalities, capable of determining whether a modality should learn from the task or its teacher. Empirical experiments on various tasks including multimodal classification, visual question answering and visual entailment demonstrate that our proposed I^{2}MCL is able to tackle the under-optimized modality problem and bring consistent improvement to multimodal learning.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

multimodal learning, curriculum learning, knowledge distillation, multi-objective optimization

*Corresponding authors. BNRist is the abbreviation of Beijing National Research Center for Information Science and Technology.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

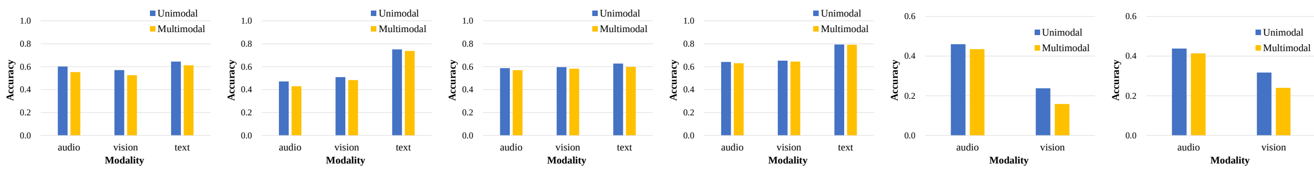
ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612468

ACM Reference Format:
Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. 2023. Intra- and Inter-Modal Curriculum for Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612468

1 INTRODUCTION

Multimodal learning is an approach to building models that can process and integrate information from multiple heterogeneous data *modalities* [6, 46], including image, text, audio, video and table. Since numerous tasks in the real world involve multiple modalities, multimodal learning has become increasingly important and attracted widespread attention as an effective way to accomplish these tasks. Existing multimodal tasks can be roughly divided into two branches based on whether the target can be predicted with a single modality or not, as stated by [46]. One is evolved from previous unimodal tasks and named *modal fusion problem*, where the modalities capable of individually predicting results are fused together to enhance predictions. Typical tasks include multimodal classification [5, 8, 11, 13, 14, 17, 29, 35–37, 43, 69, 79, 80] and regression [40–42, 58]. The other in contrast requires interaction among modalities, and in this paper, we name it *modal interaction problem*, where multiple modalities need to link, query or retrieve from each other to jointly predict the results. Typical instances include cross-modal question answering [1, 4, 25], grounding [77], reasoning [66], entailment [74] and retrieval [47, 76].

Despite the success of multimodal learning, it is widely observed that modal encoders are under-optimized and modal representations are inferior in multimodal learning compared to their unimodal counterparts, as illustrated in Figure 1. Although this problem is reported by several recent studies [19, 34, 50, 67, 70, 73], they do not reach full agreement on its definitions and causes. For example, it is called "modality failure" in [19], "greedy nature" in [73] and "modality collapse" in [34], and it is claimed to happen because of "suppression of dominant modalities" in [50] and "different convergence rates" in [67, 70]. To sum up, the mainstream thoughts about this problem are as follows: i) modality varies in optimization, so a single strategy is insufficient; ii) modality varies in dominance, so weak ones are suppressed and fail to get enough training from the task. Therefore, these existing works are devoted to balancing learning among modalities to deal with the problem.



(a) MUStARD (b) CMU-MOSI (c) UR-FUNNY (d) CMU-MOSEI (e) VGGSound [19] (f) VGGSound [50]

Figure 1: Performances of modalities in unimodal and multimodal models evaluated by linear probing, i.e., fixing trained encoder and finetuning linear output layer. Figures (a) to (d) come from our experiments, (e) from [19], and (f) from [50]. All modalities are under-optimized in multimodal learning compared to unimodal learning.

However, existing investigations and solutions to this problem suffer from two limitations. i) They only consider the inter-modal imbalance but ignore the influence of intra-modal data on each modality. An intuitive example is that when a modal encoder continuously encounters difficult or even noisy data instances beyond its competence, the representations it outputs can mislead the final result, resulting in its weak positions among all modalities. ii) They only focus on the suppression of strong modalities over weak ones, but neglect the influence of weak ones on strong ones. As long as there is a fusion or interaction module in the model, while being suppressed, the under-optimized weak ones can also be detrimental to the performance of strong ones, resulting in overall suboptimal. We will give a theoretical interpretation of this statement in Section 3.2. Therefore, it is necessary to enhance all modalities, regardless of weak or strong, from the perspective of intra-modal data to comprehensively improve multimodal learning.

Based on the foregoing discussions, in this paper, we propose to simultaneously consider intra-modal data difficulty and inter-modal modality balance. Nevertheless, how to measure data difficulty and modality imbalance remains a challenge that current works fail to tackle. Current metrics for modality imbalance include unimodal outputs, losses and performances, which can only be derived in modal fusion tasks, where outcomes from a single modality can perform predictions. But in modal interaction tasks such as visual question answering, where the answers are jointly determined by images and questions, unimodal outputs are biased or invalid, thus making the existing metrics and methods become suboptimal.

To tackle the problem and the challenge mentioned above, we propose an **I**ntra- and **I**nter-**M**odal **C**urriculum **L**earning framework (I^2MCL) for multimodal learning. In the intra-modal curriculum, we employ a pretrained teacher model for each modality to acquire knowledge distillation loss as the measurer of data difficulty, which determines the data weights within the corresponding modality, so that all modalities can be optimized in an easy-to-hard manner. In the inter-modal curriculum, we utilize a Pareto optimization strategy to measure the gradient relationship between distillation loss and task loss backpropagated to each modal encoder, which is compared across modalities to decide whether a modality should learn from the task or its teacher. As such, weak modalities can first benefit from the extra knowledge from their teachers to catch up with strong ones and then try to learn from the task, instead of learning little under the suppression all the time.

To verify the effectiveness as well as the generality of I^2MCL, we apply it to six multimodal datasets, covering both branches of multimodal tasks. The comparative empirical results with existing works demonstrate that our method can bring more improvement to multimodal learning, and the ablative experimental results present how our method works to alleviate the under-optimized modality problem. To summarize, our contributions are listed as follows,

- We present a new perspective from intra-modal data and inter-modal mutual influence to explain the under-optimized modality problem in multimodal learning.
- We propose an intra- and inter-modal curriculum framework to address the problem by considering data difficulty and modality balance, applicable to both modal fusion and interaction tasks.
- Empirical experiments demonstrate the benefit and improvement our method brings compared to existing works.

2 RELATED WORK

2.1 Multimodal Learning

Multimodal learning serves as an effective way to cope with real-world tasks involving multiple sources of data [6]. One of the earliest relevant studies is audio-visual speech recognition [78]. In the current era of deep learning and large model, there are lots of other applications such as detection, search, recommendation and generation [84] with the input of image, text, audio, video and table [62]. The taxonomy of multimodal learning is diverse. For example, it can be divided according to fusion strategies [53], model frameworks [27], multimodal challenges [6], etc. In this paper, we follow [46] and partition it into fusion problems and interaction problems on the basis of whether a single modal can make a prediction individually, so that we can clarify the universality of our method.

2.2 Under-Optimized Modality Problem

In spite of the wide application of multimodal learning, many recent works have reported the phenomenon that modalities in multimodal learning are not fully trained, optimized or exploited. It is mainly discovered in modal fusion tasks where the performances of single modalities are easy to derive. Wang et al. [70] point out unimodal networks can perform better than multimodal ones, give an explanation of different fitting rates among modalities, and propose a Gradient-Blending (GB) method. Du et al. [19] name the problem modality failure, give a hypothesis of modality imbalance and implicit bias, and propose a Uni-Modal Teacher (UMT) method. Sun et al. [67] propose a balanced learning rates method based on Adaptive Tracking Factor (ATF). Javaloy et al. [34] name the problem modality collapse and propose an impartial optimization method to mitigate it in multimodal VAEs. Wu et al. [73] come up with a greedy learner hypothesis and propose a re-balancing method based on Conditional Learning Speed (CLS). Peng et al. [50] give an opinion of modality dominance and propose an On-the-fly Gradient Modulation with Generalization Enhancement (OGM-GE) method. The most related work to ours is UMT, which also utilizes knowledge distillation to assist multimodal learning, but like all the works mentioned above, it only considers inter-modal balance and is limited in applications as we have mentioned in Section 1.

2.3 Curriculum Learning

Curriculum learning is a machine learning strategy that trains a model from easy to hard, mimicking the way that humans learn with curricula [7, 64, 71]. It can guide and denoise the machine learning process, thereby accelerating model convergence and improving model generalization. Bengio et al. [7] first give its formal definition and propose a simple method named Baby Step [65]. After that, various methods have been continuously emerging, including Self-Paced Learning [12, 21, 39], Transfer Teacher [28, 72], Reinforcement Learning Teacher [26, 83] and others [54, 59, 63]. The key components of curriculum learning include a difficulty measurer to tell what is hard to learn and a learning scheduler to decide when to learn the harder part. In this paper, we introduce our method precisely through these two components of curriculum learning for the sake of clarity.

2.4 Knowledge Distillation

Knowledge distillation refers to the transfer of knowledge from teacher models to student ones. Since it is proposed for the goal of model compression [10], teacher models are usually large-scale, ensembled and pretrained, while students are relatively small and fast, which is named offline distillation [23, 32, 57, 68, 81]. There are also methods of online distillation [3, 82], where teacher and student models are trained simultaneously, and self-distillation [15, 45], where teacher and student models are the same. Apart from the training strategy, the form of knowledge is another important component [24], which can be categorized into response-based [32], feature-based [56] and relation-based [75]. In this paper, we adopt offline distillation and feature-based knowledge.

2.5 Multi-Objective Optimization

Multi-objective optimization aims to handle the optimization problem of multiple possibly contrasting objectives [20, 49]. It is widely applied in machine learning tasks, such as multi-agent learning [51], kernel learning [44], sequential decision making [55], Bayesian optimization [31], multi-task learning [61], etc. In this paper, we use the gradient-based Pareto optimization method named multiple gradient descent algorithm (MGDA) [18, 22, 60] not to resolve the gradient conflicts but to measure and compare the relationship among the conflicting gradients across modalities as the difficulty measurer for the inter-modal curriculum. Besides, to decrease time complexity and avoid the computational bottleneck of repeat back-propagation from loss to encoders, we follow MGDA-UB proposed in [61] and approximate the gradient of a modal encoder with that of a modal representation when implementing the algorithm.

3 PRELIMINARY

3.1 Multimodal Learning

For simplicity of description, we first formulate a general definition of multimodal learning. Given a dataset $\mathcal{D} = \{(x_{i1}, ..., x_{iM}, y_i)\}_{i=1}^N$ with N data samples and M modalities, the $m^{th}(1 \le m \le M)$ modality of the $i^{th}(1 \le i \le N)$ data x_{im} can be a static one or a temporal one $x_{im} = (x_{im}^1, ..., x_{im}^T)$ of T length, and the label y_i can refer to a class, a matching, an answering, etc., according to the target task. A multimodal model aims to predict the results:

$$\hat{y}_i = f_0(f_1(x_{i1}; \theta_1), ..., f_M(x_{iM}; \theta_M); \theta_0), \tag{1}$$

where f_0 is a multimodal module parameterized by θ_0 and f_m is a unimodal encoder with θ_m. Like other machine learning tasks, the training objective is to minimize the empirical risk between predictions and truths:

$$\min_{\theta_0, \theta_1, ..., \theta_M} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{y}_i(x_{i1}, ..., x_{iM}; \theta_0, \theta_1, ..., \theta_M), y_i). \tag{2}$$

3.2 Under-Optimized Modality Problem

Modality Imbalance in Naive Fusion Settings. Literature [19] and [50] have described in detail the problem of modality imbalance in naive fusion settings. Without loss of generality, we consider the simplest multimodal model with two modalities, a concentration fusion and the optimization objective of cross-entropy loss. If we denote unimodal representations $f_m(x_{im}; \theta_m)$ as z_{im}, the multimodal representation $f_0(z_{i1}, ..., z_{iM}; \theta_0)$ as z_{i0} and loss $\mathcal{L}(\hat{y}_i, y_i)$ as l_i, the conditions above can be formulated as $M = 2, z_{i0} = W_1 z_{i1} + W_2 z_{i2} + b$, $\hat{y}_i = \text{Softmax}(z_{i0})$, and then the gradients from loss to the two modal encoders are:

$$\begin{aligned} \nabla_{\theta_1} l_i &= \nabla_{\theta_1} z_{i0} \; \nabla_{z_{i0}} l_i, \\ \nabla_{\theta_2} l_i &= \nabla_{\theta_2} z_{i0} \; \nabla_{z_{i0}} l_i, \\ \nabla_{z_{i0}} l_i &= \text{Softmax}(W_1 z_{i1} + W_2 z_{i2} + b) - y_i. \end{aligned} \tag{3}$$

If the 1^{st} modality is strong and 2^{nd} is weak, it can be concluded that the 1^{st} modality can i) dominate the gradient descent through its more contribution on $\nabla_{z_{i0}} l_i$ via $W_1 z_{i1} > W_2 z_{i2}$, because the property of Softmax is similar to Max [50]; ii) stop the optimization of the 2^{nd} modality by making $\nabla_{\theta_2} l_i \to 0$ via $\nabla_{z_{i0}} l_i \to 0$ when the 1^{st} modality has already converged [19].

Under-Optimized Modality in Multimodal Settings. Based on the explanation of modality imbalance, we further propose a theoretical analysis of the under-optimized problem due to mutual influence among all the modalities. The premise of this problem is that the multimodal model has at least a fusion or interaction module, instead of only voting at the decision level.

Consider the gradient from l_i to θ_m:

$$\nabla_{\theta_m} l_i = \nabla_{\theta_m} z_{im} \; \nabla_{z_{im}} z_{i0} \; \nabla_{z_{i0}} l_i, \tag{4}$$

and the functional relationship between z_{im} and z_{i0}:

$$z_{i0} = f_0(z_{i1}, ..., z_{iM}; \theta_0), \tag{5}$$

there must exist the function g and functions h_m such that:

$$\nabla_{z_{i0}} l_i = g(z_{i0}(z_{i1}, ..., z_{iM})), \tag{6a}$$

$$\nabla_{z_{im}} z_{i0} = h_m(z_{i1}, ..., z_{iM}, \theta_0), \tag{6b}$$

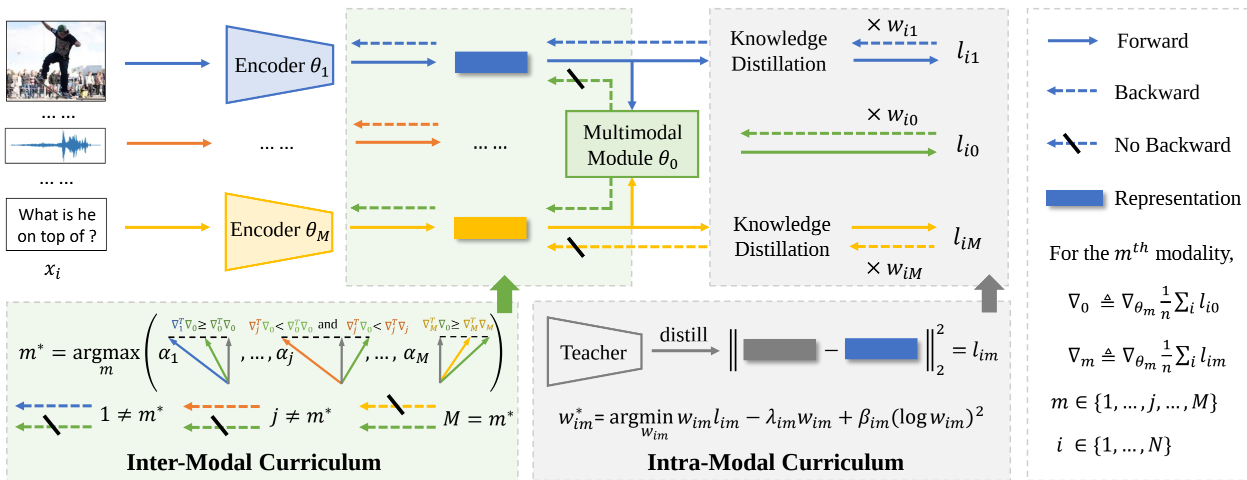


Figure 2: The overall framework of I^{2}MCL.

so the gradient backward from loss to each encoder $\nabla_{\theta_m} l_i$ are influenced by other modal representations $z_{m'} (m' \neq m)$. It is worth noting that Equation (6b) can degenerate to $\nabla_{z_{im}} z_{i0} = h_m(\theta_0)$ in naive fusion settings but Equation (6a) is always the case, e.g., $\nabla_{z_{i0}} l_i = \text{Softmax}(z_{i0}) - y_i$ for cross-entropy loss and $\nabla_{z_{i0}} l_i = z_{i0} - y_i$ for mean-squared-error loss. Therefore, the under-optimized modalities can do harm to others with their inferior representations, leading to overall suboptimal.

4 METHOD

In this section, we propose our I^{2}MCL, a generic multimodal curriculum learning method, composed of an intra-modal curriculum (Section 4.1) and an inter-modal curriculum (Section 4.2). For a clearer description, we present both of them in terms of curriculum design, difficulty measurer and learning scheduler from the perspective of curriculum learning. Lastly, we summarize the overall framework of I^{2}MCL (Section 4.3).

4.1 Intra-Modal Curriculum

Curriculum Design. The first and key step to deal with the under-optimized modality problem is to optimize them further based on their intra-modal data. We design an intra-modal curriculum by introducing offline distillation with feature-based knowledge, from which the distillation loss acts as the data difficulty measurer. It does not depend on extra labels or unimodal performances but only needs one forward propagation in each training step, ensuring the generality and efficiency of our method.

Difficulty Measurer. We employ a pretrained teacher model g_m parameterized by ϕ_m and adopt the L_2 norm distance as the optimization objective to distill knowledge for the m^{th} modality:

$$\min_{\theta_m} \frac{1}{N} \sum_{i=1}^{N} \| f_m(x_{im}; \theta_m) - g_m(x_{im}; \phi_m) \|_2^2. \tag{7}$$

We denote the distillation loss from Equation (7) as $\mathcal{L}_m$, and regard the distillation loss l_{im} of the data instance x_{im} such that $\mathcal{L}_m = \frac{1}{N} \sum_{i=1}^{N} l_{im}$ as the difficulty measurer of x_{im} for the modal encoder f_m. For the sake of comprehensiveness, we treat multimodality as a special modality, denote the task loss from Equation (2) as $\mathcal{L}_0$ and view the task loss l_{i0} of the data pair x_i as the difficulty measurer of x_i for the whole model. Within each modality, we split all data into two parts, the hard ones and the easy ones, by comparing their losses to the moving average λ_m of losses $\mathcal{L}_m$:

$$\lambda_m^{(t)} = \gamma_m \lambda_m^{(t-1)} + (1 - \gamma_m) \mathcal{L}_m^{(t)}, \tag{8}$$

where $\gamma_m \in [0, 1]$ is a discount factor and t refers to training steps. A relatively large l_{im} satisfying $l_{im}^{(t)} > \lambda_m^{(t)}$ means that the encoder has not been able to represent x_{im} well in the current training step, so we can treat x_{im} as hard data and decrease its weight, while a small loss value $l_{im}^{(t)} \leq \lambda_m^{(t)}$ represents easy data worthy of an increased weight. Based on the losses as the data difficulty measurers, we can realize specific curricula for all modalities and teach them from easy to hard with the learning scheduler described in the next paragraph.

Learning Scheduler. The scheduler that guides the learning of each modality is implemented through data reweighting. The weight w_{im} assigned to data x_{im} impacts the learning process by scaling the loss: $w_{im} l_{im}$, where w_{im} should at least satisfies the following conditions to be consistent of the core idea of curriculum learning, i.e., from easy to hard:

$$w_{im} = w_{im}(l_{im}, \lambda_m), \tag{9}$$

$$\forall\ l_{im} \leq l_{jm}, \quad \text{s.t. } w_{im} \geq w_{jm} \geq 0. \tag{10}$$

Specifically, we follow [12] and obtain w_{im} by minimizing the reweighted loss attached with curriculum regularizers, composed of a negative L_1 regularizer with λ_m to distinguish between hard and easy data, and a positive log-L_2 regularizer to force w_{im} close to 1 and avoid very high value just like weight decay :

$$\min_{w_{im}} w_{im} l_{im} - \lambda_m w_{im} + \beta_m (\log w_{im})^2, \quad (11)$$

where $\beta_m \geq 0$ is a hyperparamter to control the latter regularizer.

By treating l_{im} and λ_m as fixed values and w_{im} as a variable, the result of Equation (11) can be solved as the root of the derivative:

$$w_{im} = \begin{cases} e, & l_{im} - \lambda_m \leq -2\beta_m/e, \\ e^{-\mathcal{W}(\frac{l_{im}-\lambda_m}{2\beta_m})}, & l_{im} - \lambda_m > -2\beta_m/e, \end{cases} \quad (12)$$

where $\mathcal{W}$ refers to Lambert W function. The detailed process of deriving w_{im} value from Equation (12) with the Alternative Optimization Strategy (AOS) is described in the Appendix. Besides, the function between w_{im} and $l_{im} - \lambda_m$ is also plotted in the Appendix, visually demonstrating that w_{im} satisfies the conditions of Equation (9) and (10).

4.2 Inter-Modal Curriculum

Curriculum Design. The other indispensable step is to balance learning among modalities, preventing weak modalities from being suppressed. Therefore, we design an inter-modal curriculum to instruct modalities whether to learn from the target task or from knowledge distillation, according to the gradient relationship between distillation loss and task loss as the difficulty measurer.

Difficulty Measurer. For each modal encoder f_m, there are two gradients backward to it, i.e., $\nabla_{\theta_m}\mathcal{L}_m$ and $\nabla_{\theta_m}\mathcal{L}_0$ from distillation loss and task loss respectively. Although both of them are intended to improve the semantic representation capability of the encoder, they will inevitably conflict with each other to a certain extent. We make use of the conflict relationship as a measurer to compare which is more difficult for f_m to learn, the target task or the knowledge distillation.

Borrowing the idea of multi-objective optimization, when shared parameters are optimized by multiple objectives, the goal becomes reaching Pareto optimality by coordinating possibly contrasting directions, so we consider the optimization problem for each modal encoder with the condition of $\alpha_m \in [0, 1]$:

$$\min_{\alpha_m} \|\alpha_m \nabla_{\theta_m}\mathcal{L}_m + (1-\alpha_m)\nabla_{\theta_m}\mathcal{L}_0\|_2^2. \quad (13)$$

It is proved that the solution is either 0 or provides a direction to optimize both of them. If we abbreviate $\nabla_{\theta_m}\mathcal{L}_m$ and $\nabla_{\theta_m}\mathcal{L}_0$ as ∇_m and ∇_0 respectively for the m^{th} modality, the solution of Equation (13) can be written as:

$$\alpha_m = \begin{cases} 0, & \nabla_m^T \nabla_0 \geq \nabla_0^T \nabla_0, \\ 1, & \nabla_m^T \nabla_0 \geq \nabla_m^T \nabla_m, \\ \dfrac{(\nabla_0 - \nabla_m)^T \nabla_0}{\|\nabla_0 - \nabla_m\|_2^2}, & \text{others.} \end{cases} \quad (14)$$

The proof and calculation process along with their visualizations are included in the Appendix.

From Equation (14), α_m well reflects the magnitude and direction relationship between two gradients. For example, $\alpha_m \to 1$ means $\|\nabla_0\|_2^2 > \|\nabla_m\|_2^2$ when they form an acute angle or $\|\nabla_0\|_2^2 \gg \|\nabla_m\|_2^2$ when an obtuse angle, in which cases the m^{th} modality learns much more from task loss than distillation loss, and in other words, learning from the task is easier than from its teacher. Therefore, we take α as the difficulty measurer to decide whether to learn from the target task or from knowledge distillation.

Another point worth noticing is that procedure of resolving α_m is time-consuming especially when the number of the encoder parameters θ_m is extremely large, because it requires twice back-propagations, from $\mathcal{L}_m$ and $\mathcal{L}_0$ to θ_m respectively. To avoid this computational bottleneck, we follow [61] and calculate α_m with the estimation of $\nabla_{\theta_m}\mathcal{L}_m$ and $\nabla_{\theta_m}\mathcal{L}_0$:

$$\begin{aligned} \|\nabla_{\theta_m}\mathcal{L}_m\|_2^2 &\leq \|\nabla_{\theta_m} Z_m\|_2^2 \|\nabla_{Z_m}\mathcal{L}_m\|_2^2, \\ \|\nabla_{\theta_m}\mathcal{L}_0\|_2^2 &\leq \|\nabla_{\theta_m} Z_m\|_2^2 \|\nabla_{Z_m}\mathcal{L}_0\|_2^2, \end{aligned} \quad (15)$$

through the chain rule of gradients and the modal representations $\mathcal{Z}_m = (z_{1m}, ..., z_{Nm})$, where $z_{im} = f_m(x_{im}; \theta_m)$. Since $\nabla_{\theta_m} Z_m$ is not directly related to α_m, Equation (13) becomes:

$$\min_{\alpha_m} \|\alpha_m \nabla_{\mathcal{Z}_m}\mathcal{L}_m + (1-\alpha_m)\nabla_{\mathcal{Z}_m}\mathcal{L}_0\|_2^2, \quad (16)$$

and Equation (14) still holds with $\nabla_0 \triangleq \nabla_{\mathcal{Z}_m}\mathcal{L}_0$ and $\nabla_m \triangleq \nabla_{\mathcal{Z}_m}\mathcal{L}_m$. The approximation can significantly reduce computation time by not computing the gradients of encoders.

Learning Scheduler. In this part, we balance the learning process among modalities by guiding one modality to learn from target task loss $\mathcal{L}_0$ and others from knowledge distillation loss $\mathcal{L}_m$ based on the comparison of α_m across modalities.

At every training step, we pick the m^* modality with the largest α_m value:

$$m^* = \arg\max_m \alpha_m. \quad (17)$$

Since a large value of α_m means learning more from the target task than from knowledge distillation, the largest α_{m^*} represents the strong modality that should learn from the task first:

$$\theta_{m^*}^{(t+1)} = \theta_{m^*}^{(t)} - \eta \nabla_{\theta_{m^*}^{(t)}} \mathcal{L}_0^{(t)}, \quad (18)$$

For other modalities, we let them learn from their teachers:

$$\theta_{m'}^{(t+1)} = \theta_{m'}^{(t)} - \eta \nabla_{\theta_{m'}^{(t)}} \mathcal{L}_{m'}^{(t)}, \quad m' \neq m^*, \quad (19)$$

where η refers to the learning rate.

We design such a learning scheduler for two reasons. The first is to avoid modality suppression caused by modality imbalance. As stated in Section 1 and 3.2, weak modalities learn less than strong ones from the task, so we only let the strongest modality, i.e., the m^* one, learn from the task and instruct others to learn from teachers. As training progresses, weak modalities enhanced by teacher knowledge have the opportunity to catch up with or even become the strongest one and thus learn much from the task, so that we can keep the dynamic balance among modalities. The second reason is to avoid gradient conflicts on each modal encoder between task loss and distillation loss in the same training step. As stated in the theory of multi-objective optimization, the parameters optimized by multiple losses are likely to encounter gradient conflicts, so we force each modality to learn from only one source, either the task or the teacher, in one training step.

4.3 Multimodal Curriculum Learning

Integrating all of the above, we summarize the complete process of our I²MCL method in this subsection. It is illustrated in Figure 2, elaborated in Algorithm 1 and formulated in the Equation below:

$$\theta_j^{(t+1)} = \begin{cases} \theta_j^{(t)} - \eta \frac{1}{n} \nabla_{\theta_j^{(t)}} \sum_{i=1}^{n} w_{i0}^{(t)} l_{i0}^{(t)}, \ j \in \left\{0, \arg\max_m \alpha_m^{(t)}\right\}, \\ \theta_j^{(t)} - \eta \frac{1}{n} \nabla_{\theta_j^{(t)}} \sum_{i=1}^{n} w_{ij}^{(t)} l_{ij}^{(t)}, \ \text{others}, \end{cases} \tag{20}$$

where n is the size of minibatch. At the t^{th} step, the multimodal module and the m^* modality with the largest α_m learn from the task, other modalities learn from their teachers, and all of them learn from data in an easy-to-hard manner by data reweighting.

Concretely, we first derive the feature-based knowledge from pretrained teachers on training set data with offline computation. Then, we calculate multimodal task loss and unimodal distillation losses with once forward propagation. After that, we conduct intra- and inter-modal curriculum by updating the measurers w and α, which only add once extra backpropagation from task loss to the multimodal module and $2M$ times one-layer backpropagations to modal representations, i.e., M times from multimodal module and M times from distillation loss. It is worth noting that the backpropagations to modal representations are much more time-efficient compared with those to modal encoders. We attach the running time of our method in the Appendix to demonstrate it. Finally, we update the learnable parameters, select the best checkpoint on validation set and evaluate the final performance on test set.

5 EXPERIMENTS

In this section, we introduce the experimental setup (Section 5.1), present the performances of our method on both modal fusion tasks (Section 5.2) and modal interaction tasks (Section 5.3), and provide some further empirical analysis (Section 5.4 and 5.5).

5.1 Experimental Setup

Tasks and Datasets. We conduct experiments on both modal fusion tasks and modal interaction tasks. In modal fusion tasks, we adopt four datasets from the area of multimodal affect computing, which are provided and processed by *MultiBench* [46], a multimodal benchmark with a diverse set of datasets and algorithms for fusion problems. Following *MultiBench*, we treat these tasks as regression ones when training but as classification ones with labels of positive and negative sentiment when testing.

- **MUStARD** [13]: A dataset for multimodal sarcasm discovery, compiled from popular TV shows, consisting of audio-visual utterances annotated with sarcasm labels.
- **CMU-MOSI** [79]: A dataset for affect recognition, with a collection of video blogs from YouTube and rigorous annotation with labels for sentiment intensity in $[-3, +3]$.
- **UR-FUNNY** [29]: A dataset for multimodal humor detection, consisting of video samples from TED talks annotated with positive or negative labels.
- **CMU-MOSEI** [80]: A large dataset for emotion recognition, containing videos from YouTube and annotations of 9 discrete emotions and 3-dimensional continuous emotions.

In modal interaction tasks, we use the prevalent vision-language datasets, whose labels are determined jointly by both vision and language modalities.

- **SNLI-VE** [74]: A dataset developed from SNLI [9] and Flickr30K [76], consisting of image-sentence pairs and their relations as labels, including entailment, neutral or contradictory. We follow [74] and consider it as a three-way classification task.
- **VQA-v2** [25]: The second version of the VQA dataset [4] that builds from COCO [47] and contains open-ended questions for images, which require an understanding of vision and language to answer. We follow [2], treat it as a classification task with 3129 labels, and report the overall accuracy on test-dev set.

Comparable Methods. We compare our method with the following SOTA methods proposed for the under-optimized modality problem. The comparison experiments are mainly conducted on modal fusion tasks, for which these methods are proposed.

- **Gradient-Blending (GB)** [70]: A method to blend the gradients across modalities with weighted unimodal losses based on the overfitting-to-generalization ratios of the modalities.
- **Uni-Modal Teacher (UMT)** [19]: A method to introduce unimodal teachers to distill knowledge for all modalities.
- **Adaptive Tracking Factor (ATF)** [67]: A method to adjust the learning rates of modalities based on their unimodal losses.
- **Conditional Learning Speed (CLS)** [73]: A method to take rebalance training steps for the weak modality according to their conditional learning speed measured by unimodal performances.
- **On-the-fly Gradient Modulation with Generalization Enhancement (OGM-GE)** [50]: A method to balance modality with gradient modulation based on unimodal logits and avoid generalization drop by adding Gaussian noise.

Implementation details. To fairly evaluate our method, we apply the comparable methods and our I²MCL to strictly the same multimodal settings. In modal fusion tasks, we follow [46], build a late-fusion model composed of GRU [16] encoders, a Concat fusion module and an MLP output head, and adopt an AdamW [48] optimizer with 0.001 learning rate, 0.01 weight decay and 200 training epochs. The teacher encoders are the same as the students and pretrained in the same task. In model interaction tasks, we follow [2], build a late-fusion model with a ResNet18 [30] vision encoder, an LSTM [33] text encoder, a top-down attention layer for modal interaction and an MLP output head, and adopt an Adamax [38] optimizer with 0.002 learning rate, no weight decay and 30 training epochs. The teacher model is CLIP [52], which is 10 times larger than the student encoders and is able to output good vision and text representations. Besides, we adopt $\gamma_m = 0.9$ and $\beta_m = 1.0$ without further tuning. The concrete model architectures and other details are presented in the Appendix. With the settings above, we report the average and standard deviation results of 3 runs with different fixed random seeds on each dataset. The code is available at https://github.com/zhouyw16/I2MCL.

Table 1: Test accuracy (%) of different methods. "Uni" and "Mul" represent vanilla unimodal and multimodal learning respectively. "Audio", "Vision" and "Text" represent the performances of the modal encoders evaluated by linear probing, and "Fusion" refers to the results of modal fusion. The bold font denotes better performances in the multimodal setting.

		Uni	Mul	GB	UMT	ATF	CLS	OGM-GE	I^2MCL (ours)
MUStARD	Audio	$60.15_{1.02}$	$55.31_{1.49}$	$56.52_{1.02}$	$58.45_{1.49}$	$58.21_{0.90}$	$56.04_{0.91}$	$56.52_{0.34}$	$\mathbf{59.92}_{0.66}$
	Vision	$57.01_{0.34}$	$52.54_{1.08}$	$53.38_{0.67}$	$55.07_{0.71}$	$53.99_{0.37}$	$53.14_{0.46}$	$55.32_{0.81}$	$\mathbf{55.44}_{0.70}$
	Text	$64.49_{0.68}$	$61.23_{1.09}$	$63.53_{0.34}$	$\mathbf{64.01}_{0.49}$	$63.77_{0.81}$	$63.77_{0.49}$	$62.32_{0.49}$	$63.53_{0.90}$
	Fusion		$61.59_{1.45}$	$64.01_{0.91}$	$63.52_{0.68}$	$63.77_{0.59}$	$62.32_{0.34}$	$62.14_{1.30}$	$\mathbf{65.22}_{0.91}$
CMU-MOSI	Audio	$47.10_{1.26}$	$42.94_{0.92}$	$\mathbf{50.56}_{0.83}$	$47.79_{0.69}$	$48.63_{1.80}$	$43.26_{0.75}$	$44.67_{1.86}$	$50.15_{1.55}$
	Vision	$50.91_{0.61}$	$48.33_{1.53}$	$51.17_{0.52}$	$52.40_{0.38}$	$51.83_{1.07}$	$49.31_{1.24}$	$50.71_{0.14}$	$\mathbf{52.64}_{0.96}$
	Text	$75.10_{0.28}$	$73.78_{0.38}$	$74.54_{0.75}$	$73.55_{0.38}$	$74.04_{0.40}$	$74.32_{0.76}$	$74.09_{0.37}$	$\mathbf{75.15}_{0.12}$
	Fusion		$71.80_{1.38}$	$74.49_{0.57}$	$73.12_{0.40}$	$72.76_{0.29}$	$73.01_{1.39}$	$72.05_{1.02}$	$\mathbf{74.54}_{0.52}$
UR-FUNNY	Audio	$58.76_{0.47}$	$56.93_{1.09}$	$59.10_{0.58}$	$59.48_{0.20}$	$59.64_{0.61}$	$57.05_{0.62}$	$57.34_{0.04}$	$\mathbf{60.30}_{0.43}$
	Vision	$59.64_{0.77}$	$58.29_{0.44}$	$59.07_{0.48}$	$59.93_{0.20}$	$58.69_{0.28}$	$58.57_{0.25}$	$59.92_{0.31}$	$\mathbf{60.02}_{0.15}$
	Text	$62.76_{0.94}$	$59.80_{0.62}$	$61.09_{0.64}$	$62.79_{0.16}$	$62.57_{0.74}$	$61.75_{0.49}$	$62.56_{0.47}$	$\mathbf{62.82}_{0.25}$
	Fusion		$60.24_{1.03}$	$62.41_{0.50}$	$64.02_{0.08}$	$62.64_{0.43}$	$62.76_{0.71}$	$63.93_{0.67}$	$\mathbf{65.12}_{0.31}$
CMU-MOSEI	Audio	$64.15_{0.66}$	$62.99_{0.27}$	$63.84_{0.24}$	$64.22_{0.45}$	$63.16_{0.45}$	$63.08_{0.39}$	$63.39_{0.46}$	$\mathbf{64.45}_{0.39}$
	Vision	$65.25_{0.25}$	$64.46_{0.31}$	$\mathbf{65.73}_{0.65}$	$65.64_{0.22}$	$65.54_{1.53}$	$64.45_{0.23}$	$64.72_{0.60}$	$65.39_{0.50}$
	Text	$79.34_{0.18}$	$79.15_{0.28}$	$79.52_{0.21}$	$79.21_{0.16}$	$79.15_{0.50}$	$79.11_{0.64}$	$79.22_{0.09}$	$\mathbf{79.54}_{0.37}$
	Fusion		$80.20_{0.60}$	$80.60_{0.55}$	$80.49_{0.07}$	$80.20_{0.48}$	$80.26_{0.51}$	$80.25_{0.54}$	$\mathbf{81.05}_{0.47}$

5.2 Results of Modal Fusion Tasks

Table 1 reports the comparison with existing methods over four datasets in terms of test accuracy for binary classification. It is shown that the proposed I^2MCL method can outperform all the multimodal baselines consistently on both individual modalities and multimodal fusion. Specifically, we have the following observations. i) Compared with the methods only adjusting the size of loss, gradient and learning rate, like ATF, CLS and OMG-GE, our I^2MCL incorporates knowledge from pretrained teachers and achieves relatively large improvement. ii) Compared with the methods with additional losses, like GB and UMT, our I^2MCL is carefully designed with a two-level curriculum to guide how the modalities learn from data and losses, and thus outperforming them substantially. iii) Our I^2MCL can even outperform unimodal learning on some modalities over the datasets like CMU-MOSI, UR-FUNNY and CMU-MOSEI, which is mainly due to the combined effect of knowledge distillation, curriculum learning and mutual promotion among modalities in multimodal learning.

Table 2: Test-dev accuracy on VQA-v2 and test accuracy on SNLI-VE. "KD" means adding knowledge distillation loss to optimization objective; "Intra" means adding intra-modal curriculum; "Inter" means adding inter-modal curriculum.

KD	Intra	Inter	VQA-v2	SNLI-VE
			$51.77_{0.19}$	$68.83_{0.07}$
✓			$53.16_{0.19}$	$68.91_{0.15}$
✓	✓		$53.64_{0.15}$	$69.11_{0.14}$
✓		✓	$54.14_{0.20}$	$70.63_{0.08}$
✓	✓	✓	$\mathbf{54.34}_{0.11}$	$\mathbf{70.95}_{0.13}$

5.3 Results of Modal Interaction Tasks

Table 2 presents the results over VQA-v2 and SNLI-VE datasets. Since other baselines are not proposed and not suitable for these tasks, we evaluate our method by means of the ablation experiment, which presents the effects of knowledge distillation, intra-modal curriculum and inter-modal curriculum. It is shown that the combination of these strategies can improve model performances significantly and reach average absolute improvements of 2.57% on VQA test-dev set and 2.12% on SNLI-VE test set.

5.4 Analysis of Intra-Modal Curriculum

To further analyze how our I^2MCL method works, we visualize the intra-modal curriculum process by tracking the weight changes of typical "hard" and "easy" data within different modalities in Table 3. According to the definition in Section 4.1, the weight range is $[0, e]$, and the larger the weight, the simpler the data. It can be observed that complex images or questions have smaller weights at different epochs, while simple ones always possess larger weights.

5.5 Analysis of Inter-Modal Curriculum

Figure 3 illustrates the process of inter-modal curriculum, from which we have the following observations. i) All modalities have the opportunity to learn from the task in the training process, ensuring all of them can learn the task-specific knowledge. ii) The text modality has relatively large α and it learns from the task more often, which is consistent with its better performance than other modalities as shown in Table 1, verifying that it is reasonable to choose α as the inter-modal measurer. iii) The α values of weak modalities like vision and audio generally increase with training steps thanks to the knowledge they learn from teachers, which phenomenon reflects the overall improvement of modality balance.

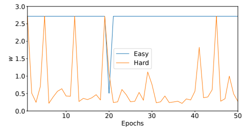
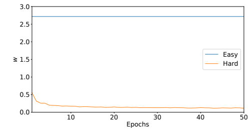
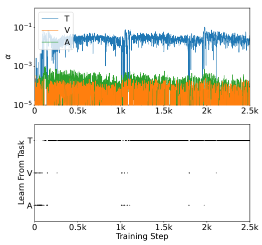
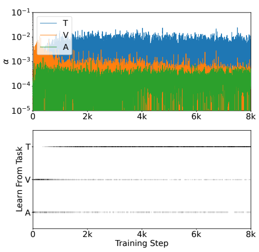
Easy Image	Hard Image	Weights	Easy Question	Hard Question	Weights
			Is the water calm?	What infrastructure would one drive on to go perpendicular to the water?	

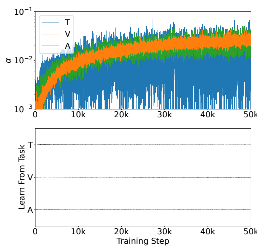
Table 3: The visualization of intra-modal curriculum with VQA-v2 data.



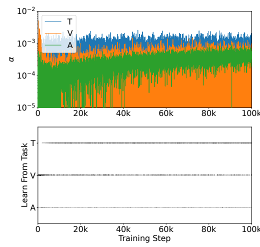
(a) MUStARD



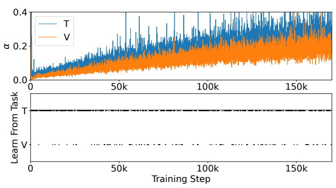
(b) CMU-MOSI



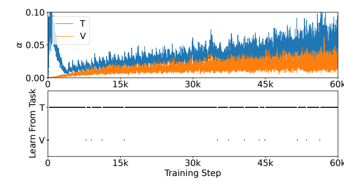
(c) UR-FUNNY



(d) CMU-MOSEI



(e) VQA-v2



(f) SNLI-VE

Figure 3: The visualization of inter-modal curriculum. For each figure, the upper subfigure illustrates how the inter-modal criterion α_m changes with training steps for each modality, and the lower one illustrates which modality is the one with the largest α_m, learning from the task. "T", "V" and "A" refer to Text, Vision and Audio modality respectively.

Besides, we compare our inter-modal curriculum with some other possible strategies where vision and text modalities learn from the task and teachers in different ways, and the results are reported in Table 4. It is observed that our I^{2}MCL in the last row outperforms all other strategies. i) Compared with the first four rows where each modality can only learn from one source, I^{2}MCL can guide them dynamically to learn from both sources. ii) Compared with the fifth row, I^{2}MCL avoids the possible gradient conflict between two losses in each training step. iii) Compared with the sixth and seventh row, I^{2}MCL comprehensively considers α across modalities and schedules the learning according to its relative size instead of its absolute size.

6 CONCLUSION

In this paper, we point out the under-optimized modality problem in multimodal learning from a new perspective of intra-modal data and inter-modal mutual influence, based on which we propose I^{2}MCL, a multimodal learning method with intra- and inter-modal curriculum considering both data difficulty and modality balance to address the issue. The method is generic enough to be applied to various multimodal settings, covering both modal fusion and interaction tasks. A possible and promising future direction is to adapt it to the pretraining or finetuning process of large multimodal models in this era of deep learning and large models.

Table 4: Results of possible inter-modal curriculum strategies. The values in the first four columns represent how much to learn from task or teacher. $\mathbb{I}$ is an indicator function.

Vision		Text		VQA-v2	SNLI-VE
Task	Teacher	Task	Teacher	+KD+Intra	+KD+Intra
1	0	1	0	$51.85_{0.18}$	$68.86_{0.09}$
0	1	0	1	$53.86_{0.16}$	$67.73_{0.06}$
1	0	0	1	$50.60_{0.13}$	$67.04_{0.12}$
0	1	1	0	$53.89_{0.10}$	$70.63_{0.21}$
1	1	1	1	$53.64_{0.15}$	$69.11_{0.14}$
α_v	$1-\alpha_v$	α_t	$1-\alpha_t$	$52.93_{0.13}$	$67.98_{0.09}$
$1-\alpha_v$	α_v	$1-\alpha_t$	α_t	$53.16_{0.14}$	$68.96_{0.17}$
$\mathbb{I}(\alpha_v \geq \alpha_t)$	$\mathbb{I}(\alpha_v < \alpha_t)$	$\mathbb{I}(\alpha_v \leq \alpha_t)$	$\mathbb{I}(\alpha_v > \alpha_t)$	$\mathbf{54.34}_{0.11}$	$\mathbf{70.95}_{0.13}$

ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China No. 62250008, 62222209, 62102222, Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[3] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235* (2018).

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[5] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992* (2017).

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 446–461.

[9] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[10] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 535–541.

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

[12] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. 2020. SuperLoss: A Generic Loss for Robust Curriculum Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[13] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).

[14] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

[15] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. 2021. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3119–3124.

[16] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[17] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*. 845–854.

[18] Jean-Antoine Désidéri. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique* 350, 5-6 (2012), 313–318.

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

[20] Matthias Ehrgott. 2005. *Multicriteria optimization*. Vol. 491. Springer Science & Business Media.

[21] Yanbo Fan, Ran He, Jian Liang, and Baogang Hu. 2017. Self-paced learning: An implicit regularization perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[22] Jörg Fliege and Benar Fux Svaiter. 2000. Steepest descent methods for multicriteria optimization. *Mathematical methods of operations research* 51 (2000), 479–494.

[23] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*. PMLR, 1607–1616.

[24] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129 (2021), 1789–1819.

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[26] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *international conference on machine learning*. PMLR, 1311–1320.

[27] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access* 7 (2019), 63373–63394.

[28] Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*. PMLR, 2535–2544.

[29] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618* (2019).

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[31] Daniel Hernández-Lobato, Jose Hernandez-Lobato, Amar Shah, and Ryan Adams. 2016. Predictive entropy search for multi-objective bayesian optimization. In *International conference on machine learning*. PMLR, 1492–1501.

[32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[34] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. In *International Conference on Machine Learning*. PMLR, 9938–9964.

[35] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[36] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[37] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems* 33 (2020), 2611–2624.

[38] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[39] M Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-Paced Learning for Latent Variable Models. In *NIPS*, Vol. 1. 2.

[40] Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. 2021. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 909–916.

[41] Michelle A Lee, Brent Yi, Roberto Martín-Martín, Silvio Savarese, and Jeannette Bohg. 2020. Multimodal sensor fusion with differentiable filters. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10444–10451.

[42] Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. 2020. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics* 36, 3 (2020), 582–596.

[43] Luis A Leiva, Asutosh Hota, and Antti Oulasvirta. 2020. Enrico: A dataset for topic modeling of mobile UI designs. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–4.

[44] Cong Li, Michael Georgiopoulos, and Georgios C Anagnostopoulos. 2014. Pareto-path multitask multiple kernel learning. *IEEE transactions on neural networks and learning systems* 26, 1 (2014), 51–61.

[45] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[46] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502* (2021).

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[48] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[49] Kaisa Miettinen. 1999. *Nonlinear multiobjective optimization*. Vol. 12. Springer Science & Business Media.

[50] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8238–8247.

[51] Fabrice Poirion, Quentin Mercier, and Jean-Antoine Désidéri. 2017. Descent algorithm for nonsmooth stochastic multiobjective optimization. *Computational Optimization and Applications* 68 (2017), 317–331.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[53] Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* 34, 6 (2017), 96–108.

[54] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*. PMLR, 4334–4343.

[55] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.

[56] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).

[57] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[58] Marcelo Sardelich and Suresh Manandhar. 2018. Multimodal deep learning for short-term stock volatility prediction. *arXiv preprint arXiv:1812.10479* (2018).

[59] Shreyas Saxena, Oncel Tuzel, and Dennis DeCoste. 2019. Data parameters: A new family of parameters for learning a differentiable curriculum. *Advances in Neural Information Processing Systems* 32 (2019).

[60] Stefan Schäffler, Reinhart Schultz, and Klaus Weinzierl. 2002. Stochastic method for the solution of unconstrained vector optimization problems. *Journal of Optimization Theory and Applications* 114 (2002), 209–222.

[61] Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems* 31 (2018).

[62] Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. 2021. Multimodal automl on structured tables with text fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*.

[63] Samarth Sinha, Animesh Garg, and Hugo Larochelle. 2020. Curriculum By Smoothing. *Advances in Neural Information Processing Systems* 33 (2020).

[64] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision* 130, 6 (2022), 1526–1565.

[65] Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 751–759.

[66] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491* (2018).

[67] Ya Sun, Sijie Mai, and Haifeng Hu. 2021. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters* 28 (2021), 1650–1654.

[68] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*. PMLR, 10347–10357.

[69] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[70] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multimodal classification networks hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12695–12705.

[71] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[72] Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*. PMLR, 5238–5246.

[73] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*. PMLR, 24043–24055.

[74] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706* (2019).

[75] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4133–4141.

[76] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[77] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 69–85.

[78] Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. 1989. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine* 27, 11 (1989), 65–71.

[79] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[80] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.

[81] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928* (2016).

[82] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4320–4328.

[83] Mingjun Zhao, Haijiang Wu, Di Niu, and Xiaoli Wang. 2020. Reinforced Curriculum Learning on Pre-Trained Neural Machine Translation Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9652–9659.

[84] Wenwu Zhu, Xin Wang, and Wen Gao. 2020. Multimedia intelligence: When multimedia meets artificial intelligence. *IEEE Transactions on Multimedia* 22, 7 (2020), 1823–1835.

A APPENDIX FOR METHOD

A.1 I²MCL Algorithm

Algorithm 1 I²MCL Algorithm

Require: The moving average factor γ_m, regularizer coefficient β_m and pretrained teacher model g_m parameterized by ϕ_m.
- **Initialize** the multimodal model parameters $\{\theta_0, \theta_m\}$.
- **Precompute** feature-based knowledge $g_m(x_{im}; \phi_m)$.
- **while** not convergent **do**
  - Calculate multimodal task loss l_{i0} via Eq. (2);
  - Calculate unimodal distillation loss l_{im} via Eq. (7);
  - Update w_{i0} and w_{im} for the intra-model curriculum via Eq. (8) and (12);
  - Update α_m for the inter-modal curriculum via Eq. (14);
  - Update θ_0 and θ_m via Eq. (20).
- **end while**
- **Return** $\{\theta_0^*, \theta_m^*\}$.

A.2 Intra-Modal Curriculum

A.2.1 *Calculation of Data Weights w_{i0} and w_{im}.*

In this paper, we follow Superloss [12] and obtain the optimization objective:

$$w_{im} = \arg\min_{w_{im}} w_{im} l_{im} - \lambda_m w_{im} + \beta_m (\log w_{im})^2, \tag{21}$$

The value of w_{im} can be resolved with **Alternative Optimization Strategy (AOS)**. First, we fix w_{im} and calculate the loss value l_{im} of each data x_{im} within the m^{th} modality parameterized by θ_m:

$$l_{im} = \mathcal{L}(x_{im}; \theta_m, w_{im}), \tag{22}$$

and define the difficulty criterion λ_m as the moving average of loss:

$$\lambda_m^{(t)} = \gamma_m \lambda_m^{(t-1)} + (1-\gamma_m)\frac{1}{N}\sum_{i=1}^{N} l_{im}^{(t)}, \tag{23}$$

where $\gamma_m \in [0, 1]$ is a discount factor and t refers to training steps.

Then, we fix l_{im} and λ_m to obtain w_{im} by resolving the derivative of Equation (21) with the condition of $w_{im} > 0$:

$$\begin{aligned}
&\frac{\partial}{\partial w_{im}}\left(w_{im} l_{im} - \lambda_m w_{im} + \beta_m (\log w_{im})^2\right) = 0,\\
\iff\ & (l_{im} - \lambda_m) w_{im} + 2\beta_m (\log w_{im}) = 0,\\
\iff\ & \frac{l_{im} - \lambda_m}{2\beta_m} = -\frac{\log w_{im}}{w_{im}},\\
\iff\ & c = d e^d,\\
\text{where}\quad & c = \frac{l_{im} - \lambda_m}{2\beta_m} \in \mathbb{R},\ d = -\log w_{im} \in \mathbb{R}.
\end{aligned} \tag{24}$$

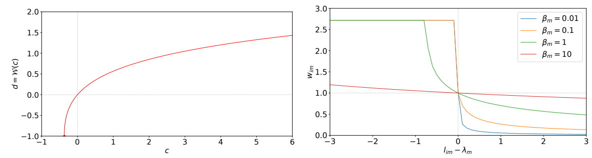
The solution is $d = \mathcal{W}(c), c \geq -\frac{1}{e}$, where $\mathcal{W}$ refers to Lambert W function. When $c \leq -\frac{1}{e}$, we define $d = \mathcal{W}(-\frac{1}{e}) = -1$ to guarantee the continuity of the function.

To sum up, we can give the solution of w_{im}:

$$w_{im} = \begin{cases} e, & l_{im} - \lambda_m \leq -2\beta_m/e, \\ e^{-\mathcal{W}\left(\frac{l_{im}-\lambda_m}{2\beta_m}\right)}, & l_{im} - \lambda_m > -2\beta_m/e, \end{cases} \tag{25}$$

A.2.2 *Visualization.*

It can be observed that when $l_{im} > \lambda_m$, $w_{im} > 1$ and when $l_{im} < \lambda_m$, $w_{im} < 1$, which encourages the model to learn more from easy data and reduces the impact of difficult data.



(a) Lambert W function. **(b) $w_{im} = w_{im}(l_{im}, \lambda_m)$.**

A.3 Inter-Modal Curriculum

A.3.1 *MGDA.*

MGDA utilizes the Karush-Kuhn-Tucker (KKT) condition. For a modal encoder f_m with θ_m, the KKT condition is that there exists an $\alpha_m \in [0, 1]$ such that $\alpha_m \nabla_m + (1-\alpha_m)\nabla_0 = 0$. The solution satisfying the condition is named a Pareto stationary point, which can be acquired by considering the optimization problem:

$$\min_{\alpha_m} \|\alpha_m \nabla_m + (1-\alpha_m)\nabla_0\|_2^2, \tag{26}$$

[18] has proved that the solution is either 0 satisfying the KKT condition, or provides a direction to guide both of the gradients. Therefore, we can measure the gradient relationship by solving Equation (26):

$$\begin{aligned}
&(\nabla_m - \nabla_0)(\alpha_m \nabla_m + (1-\alpha_m)\nabla_0) = 0,\\
\iff\ & \alpha_m \nabla_m^2 + (1-2\alpha_m)\nabla_m^T \nabla_0 - (1-\alpha_m)\nabla_0^2 = 0,\\
\iff\ & \alpha_m(\nabla_m^2 - 2\nabla_m^T\nabla_0 + \nabla_0^2) = \nabla_0^2 - \nabla_m^T\nabla_0,\\
\iff\ & \alpha_m = \frac{(\nabla_0 - \nabla_m)^T \nabla_0}{\|\nabla_0 - \nabla_m\|_2^2}.
\end{aligned} \tag{27}$$

A.3.2 *Visualization.*

Figure 5 visualizes Equation (14).

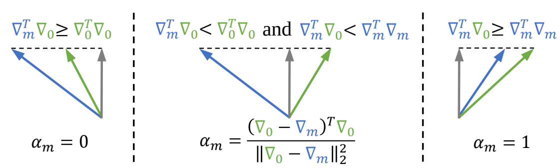


Figure 5: Visualization of α_m.

A.3.3 *Efficient MGDA.*

Furthermore, we follow [61] to avoid the time consumption on the calculation of gradients $\|\nabla_{\theta_m}\mathcal{L}_0\|$ and $\|\nabla_{\theta_m}\mathcal{L}_m\|$ by approximating them with $\|\nabla_{Z_m}\mathcal{L}_0\|$ and $\|\nabla_{Z_m}\mathcal{L}_m\|$, where Z_m is modal representations output by modal encoders.

$$\begin{aligned}
&\|\alpha_m \nabla_m + (1-\alpha_m)\nabla_0\|_2^2,\\
=\ & \|\alpha_m \nabla_{\theta_m}\mathcal{L}_m + (1-\alpha_m)\nabla_{\theta_m}\mathcal{L}_0\|_2^2,\\
\leq\ & \|\nabla_{\theta_m} Z_m\|_2^2\ \|\alpha_m \nabla_{Z_m}\mathcal{L}_m + (1-\alpha_m)\nabla_{Z_m}\mathcal{L}_0\|_2^2,
\end{aligned} \tag{28}$$

Since $\nabla_{\theta_m} Z_m$ is not directly related to α_m, we can drop $\nabla_{\theta_m} Z_m$ and optimize (29) instead of Equation (26):

$$\min_{\alpha_m} \|\alpha_m \nabla_{Z_m}\mathcal{L}_m + (1-\alpha_m)\nabla_{Z_m}\mathcal{L}_0\|_2^2. \tag{29}$$

Table 5: Dataset Information

Dataset	Task	Modality	Training Set	Validation Set	Test Set	Metrics	Classes
MUStARD	Multimodal Classification	Audio, Vision, Text	412	137	138	Accuracy	2
CMU-MOSI	Multimodal Classification	Audio, Vision, Text	1283	214	686	Accuracy	2
UR-FUNNY	Multimodal Classification	Audio, Vision, Text	8074	1034	1058	Accuracy	2
CMU-MOSEI	Multimodal Classification	Audio, Vision, Text	16265	1869	4643	Accuracy	2
VQA-v2	Visual Question Answering	Vision, Text	443757	214354	447793	Accuracy	3129
SNLI-VE	Visual Entailment	Vision, Text	529527	17858	17901	Accuracy	3

Table 6: Test accuracy (%). "KD" means adding knowledge distillation loss to optimization objective; "Intra" means adding intra-modal curriculum; "Inter" means adding inter-modal curriculum.

KD	Intra	Inter	MUStARD				CMU-MOSI				UR-FUNNY				CMU-MOSEI			
			Audio	Vision	Text	Fusion	Audio	Vision	Text	Fusion	Audio	Vision	Text	Fusion	Audio	Vision	Text	Fusion
✓			56.76	52.66	62.32	62.80	48.07	51.88	74.59	73.32	59.17	59.92	62.38	63.33	63.96	64.77	78.95	80.21
✓	✓		58.94	53.38	63.41	63.04	49.65	51.93	74.79	73.68	59.74	59.55	62.51	64.43	63.45	65.19	79.53	80.50
✓	✓	✓	**59.92**	**55.44**	**63.53**	**65.22**	**50.15**	**52.64**	**75.15**	**74.54**	**60.30**	**60.02**	**62.82**	**65.12**	**64.45**	**65.39**	**79.54**	**81.05**

Table 7: Training time (second per epoch). We run the methods multiple times on the same GPU and report the average time. "Mul" represents vanilla multimodal learning.

	Mul	**GB**	**UMT**	**ATF**	**CLS**	**OMG-GE**	**+Intra**	**+Inter**	**I^{2}MCL** w/o Eq. (16)	**I^{2}MCL** w/ Eq. (16)
CMU-MOSEI	12.09	55.48	12.28	23.69	14.97	14.81	13.63	14.73	23.17	16.55
SNLI-VE	890	-	-	-	-	-	914	936	1903	955

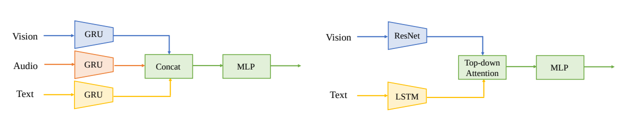
B APPENDIX FOR EXPERIMENTS

B.1 Dataset Information

Table 5 summarizes the information of all datasets involved in this paper. The first four datasets belong to the *Modal Fusion Task* and the last two are included in *Modal Interaction Task*.

B.2 Model Architecture

Figure 6 depicts the architectures of the models we build for modal fusion and interaction tasks, following *MultiBench* [1] and *Bottom-Up-Attention* [2] respectively.



(a) For modal fusion tasks. **(b) For modal interaction tasks.**

Figure 6: Model Architecture.

B.3 Compared with Pretrained Teacher

Table 8 reports the gap between learning from the teacher and directly utilizing features output by the teacher. It is observed that under the same condition of a top-down attention [2] interaction module and an MLP output head, the model trained through I^{2}MCL can outperform its counterpart with pretrained encoders on SNLI-VE test set. Although our method cannot outperform on VQA-v2 test-dev set, it is acceptable because the ResNet18 and LSTM encoders are 10 times smaller than CLIP encoders and they are not pretrained on large external data like CLIP.

Table 8: Comparison between learning from teachers (I^{2}MCL) and directly utilizing teachers' features (Pretrained Teacher).

	VQA-v2	SNLI-VE
Pretrained Teacher	$\mathbf{57.67}_{0.03}$	$70.60_{0.11}$
I^{2}MCL (Ours)	$54.34_{0.11}$	$\mathbf{70.95}_{0.13}$

B.4 Ablation Study on Modal Fusion Tasks

We conduct the ablation study on modal fusion tasks and compare the results by gradually adding knowledge distillation, intra-modal curriculum and inter-modal curriculum in Table 6. It is observed that each part of our method plays a key role on performance improvement.

B.5 Analysis of Time Complexity

We report the average time per epoch of the comparative methods and our I^{2}MCL over CMU-MOSEI and SNLI-VE in Table 7, which demonstrates that the time complexity of our I^{2}MCL is acceptable.

[1] https://github.com/pliang279/MultiBench
[2] https://github.com/hengyuan-hu/bottom-up-attention-vqa