

# Mitigating Modality Imbalance in Multi-modal Learning via Multi-objective Optimization

Heshan Fernando<sup>1</sup>Parikshit Ram<sup>2</sup>Yi Zhou<sup>2</sup>Soham Dan<sup>3†</sup>Horst Samulowitz<sup>2</sup>Nathalie Baracaldo<sup>2</sup>Tianyi Chen<sup>1,4†</sup><sup>1</sup>Rensselaer Polytechnic Institute<sup>2</sup>IBM Research<sup>3</sup>Microsoft<sup>4</sup>Cornell University

## Abstract

Multi-modal learning (MML) aims to integrate information from multiple modalities, which is expected to lead to superior performance over single-modality learning. However, recent studies have shown that MML can underperform, even compared to single-modality approaches, due to imbalanced learning across modalities. Methods have been proposed to alleviate this imbalance issue using different heuristics, which often lead to computationally intensive subroutines. In this paper, we reformulate the MML problem as a multi-objective optimization (MOO) problem that overcomes the imbalanced learning issue among modalities and propose a gradient-based algorithm to solve the modified MML problem. We provide convergence guarantees for the proposed method, and empirical evaluations on popular MML benchmarks showcasing the improved performance of the proposed method over existing balanced MML and MOO baselines, with up to  $\sim 20\times$  reduction in subroutine computation time. Our code is available at <https://github.com/heshandevaka/MIMO>.

extend MML by scaling model sizes, training on diverse multi-modal datasets, and adopting architectures capable of capturing richer cross-modal interactions. Models such as Gemini (Team et al., 2023), GPT-4 (Achiam et al., 2023), and Gato (Reed et al., 2022) have demonstrated exceptional performance across a range of downstream tasks. Unlike uni-modal methods, MML combines complementary information from various sources, leading to richer task representations. For example, combining visual and textual data can enhance understanding in vision-language tasks (Zhou et al., 2023; Oldfield et al., 2023; Wei et al., 2024).

A common MML approach uses separate encoders for each modality to transform data into feature representations that are “fused” before further processing. These fused feature representations are then processed by a model head to produce the output. In the two-modality case, MML problem can be formulated as:

$$\min_{\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}} f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}) \quad (1)$$

where  $f_{mm}$  is the multi-modal loss,  $\theta_{m_1}$  and  $\theta_{m_2}$  are encoders for modalities  $m_1$  and  $m_2$ , and  $\vartheta_{mm}$  is the multi-modal head that processes the fused features. The goal is then to optimize  $f_{mm}$  in (1) using standard optimization algorithms (e.g. SGD).

## 1.1 Imbalance issue in multi-modal learning

While MML has the potential to outperform uni-modal learning by providing richer task representations, recent studies have shown that standard MML approaches do not always improve performance compared to the best-performing uni-modal models (Wang et al., 2020; Huang et al., 2022). This highlights an inefficiency in current MML methods when it comes to effectively exploiting and integrating information from multiple modalities, presenting a significant challenge to the field. In some recent works (Ma et al., 2022; Hu et al., 2022; Peng et al., 2022), this inefficiency is sometimes attributed to the existence of a dominant modality,

## 1 Introduction

Multi-modal learning (MML) has gained attention for its ability to leverage information from multiple data modalities, such as images, text, audio, and video (Wang et al., 2022; Shridhar et al., 2020; Zhang et al., 2019). An important advancement in this field is the emergence of large multi-modal models (LMMs), which

<sup>‡</sup>The work was done when the author was at IBM.

<sup>†</sup>This work was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration.

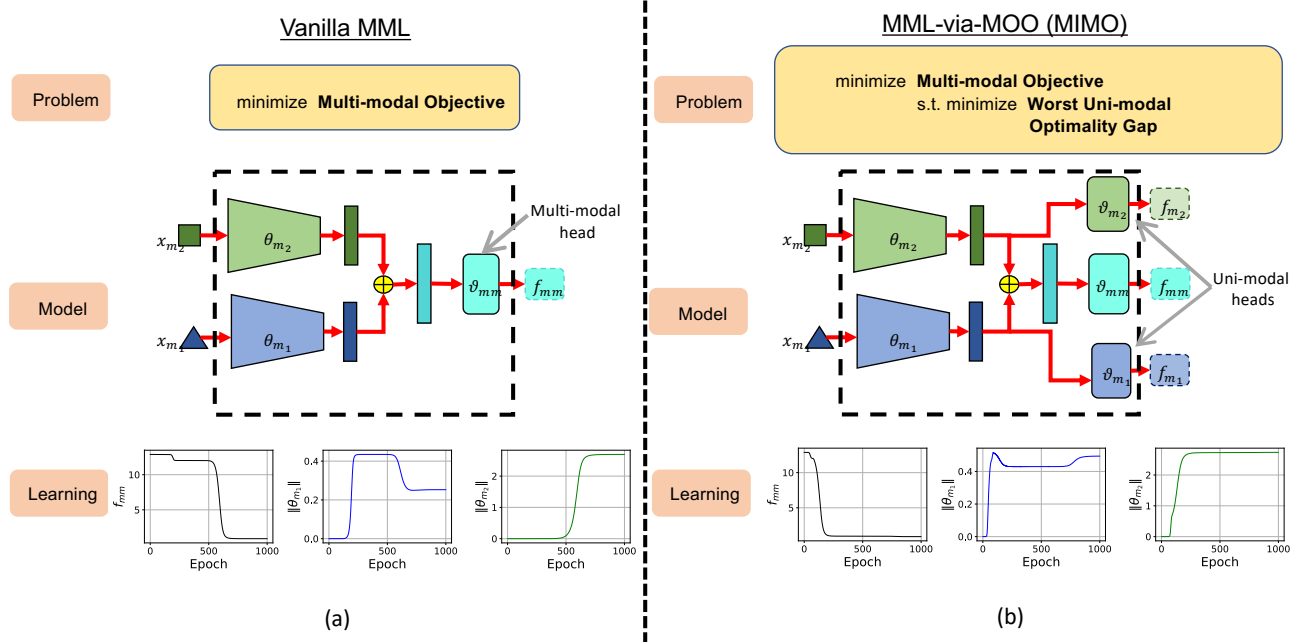


Figure 1: **Balanced multi-modal learning via multi-objective optimization.** (a): Optimizing the standard MML objective can lead to slower convergence, due to fast to learn modalities dominating the optimization process (b): We propose MML-via-MOO (MIMO), which optimizes a modified MML objective. This allows the multi-modal network to avoid dominance by one modality, which leads to faster convergence.

which prevents the model from fully exploiting the other relatively non-dominant data modalities. Using a toy example, we demonstrate this imbalance in learning different modalities in Figure 1 (a). Note how weights of  $\theta_{m_1}$  and  $\theta_{m_2}$  are learned at different speeds (See Figure 1 (a)  $\|\theta_{m_1}\|$  and  $\|\theta_{m_2}\|$  curves), leading to slow convergence for the multi-modal model.

Recent studies have theoretically explained this inefficiency by investigating the training process of late fusion models. In (Allen-Zhu and Li, 2020; Huang et al., 2022; Han et al., 2022), the emergence of the dominant modality has been explained via the concept of modality competition. In (Zhang et al., 2023b), the imbalance in modality during learning is attributed to the statistical characteristics of multi-modal datasets. In addition to the theoretical studies, empirical works attempt to develop methods to modulate the training of a multi-modal model and balance the learning of different modalities, and thus, achieve better performance (Peng et al., 2022; Fujimori et al., 2020; Yao and Mihalcea, 2022; Peng et al., 2022; Li et al., 2023; Wei and Hu, 2024) (Please refer to Appendix E for an extensive discussion on related works). However, a principled algorithm design that involves low computational complexity per iteration for modality balance in MML with convergence guarantees is missing.

## 1.2 Our contributions

In this work, we aim to address the modality imbalance issue in MML using a modified problem formulation. The key idea behind the modified formulation is to encourage the learning of slow-to-learn modalities. Intuitively, the slow-to-learn modalities have comparatively poor performance, considering the uni-modal objectives. Thus, in the proposed modified problem, we augment the multi-modal objective with modality-specific (uni-modal) objectives, which will be given preference over a multi-modal objective. This preference is enforced by optimizing the multi-modal objective under the constraint of optimality of the worst-performing uni-modal objective. More concretely, we reformulate the MML problem in (1) as a lexicographic MOO problem (Miettinen, 1999), given by

$$\begin{aligned}
 & \min_{\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}} f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}) \\
 & \text{s.t. } \theta_{m_1}, \theta_{m_2}, \vartheta_{m_1}, \vartheta_{m_2} \\
 & \quad \in \arg \min_{\theta_{m_1}, \theta_{m_2}} \max_{k \in \{1, 2\}} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*).
 \end{aligned} \tag{2}$$

where  $f_{m_k}$  for  $k \in \{1, 2\}$  are uni-modal objectives induced by separate uni-modal heads  $\vartheta_{m_1}$  and  $\vartheta_{m_2}$  (see Figure 1 (b)), and  $f_{m_k}^* = \min_{\Theta_{m_k}, \vartheta_{m_k}} f_{m_k}(\Theta_{m_k}, \vartheta_{m_k})$ ,

for all  $k \in \{1, 2\}$ . Note that the formulation in (2) can be easily extended to accommodate an arbitrary number of modalities; however, for the sake of clarity, here we mainly focus on the two-modality case. We then propose MML-via-MOO (MIMO), a simple gradient-based algorithm to solve the new formulation. As shown in Figure 1 (b), our method can alleviate the modality imbalance issue, converging fast to the global optimum. Specifically, note how weights of  $\theta_{m_1}$  and  $\theta_{m_2}$  are learned at similar speeds (See Figure 1 (b)  $\|\theta_{m_1}\|$  and  $\|\theta_{m_2}\|$  curves), leading to faster convergence. We also establish the convergence of our method theoretically. We then benchmark our algorithm against state-of-the-art MML methods, demonstrating up to a  $\times 3$  reduction in computation time while achieving superior performance.

## 2 Preliminaries

In this section, we introduce some basics in both MML (Wang et al., 2020) and MOO (Miettinen, 1999).

### 2.1 Multi-modal learning (MML)

Consider the classification problem using a multi-modal dataset  $\mathcal{D}_{mm} := \{x_i^{(m_1)}, x_i^{(m_2)}, \dots, x_i^{(m_K)}, y_i\}_{i=1}^N$ , which consists  $N$  input data  $x_i^{(m_k)}$  from  $K$  modalities  $m_k$ , where  $k \in [K] := \{1, 2, \dots, K\}$ , and the corresponding labels  $y_i$ . Unlike in the uni-modal case, we need a “fusion” strategy to fuse the inputs from different modalities together before producing the output for the loss function. Depending on whether the fusion happens during the feature extraction, after feature extraction, or in a hybrid manner, the fusion strategies can be classified as *early*, *late*, or *hybrid fusion*, respectively (Li et al., 2023). In early fusion, data from different modalities are processed together starting from raw inputs to obtain multi-modal features. In late fusion, separate encoders are used to extract uni-modal features and then fused at the final stage of the model. Any combination of early and late fusions is known as hybrid fusion. We consider the late fusion strategy, where the fusion of different modalities is done after extracting features from each modality (See Figure 1 (a)). Some common examples of fusion are summation of the extracted features (summation) or concatenating extracted features (concatenation) (Peng et al., 2022). Let  $\theta_{m_k}$  be the parameter for an encoder that extracts features from inputs  $x_i^{(m_k)}$  for all  $k \in [K]$ . Let  $\vartheta_{mm}$  be the parameter for a multi-modal head that maps the fused extracted features to target output. Then, one can formulate the problem of finding the optimal multi-modal model as

$$\min_{\Theta_{mm} \in \mathbb{R}^{d_{mm}}} f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}, \dots, \theta_{m_K}), \quad (3)$$

where  $\Theta_{mm} := [\vartheta_{mm}; \theta_{m_1}; \theta_{m_2}; \dots; \theta_{m_K}]$ , and  $f_{mm} : \mathbb{R}^{d_{mm}} \mapsto \mathbb{R}$  is the multi-modal objective that is defined by all input modalities in  $\mathcal{D}_{mm}$ . Solving (3), one can find the optimal MML model  $\Theta_{mm}^*$ .

**Uni-modal bias in MML.** As described in Section 1, the imbalance in MML leads to poor performance of MML, even compared to uni-modal learning. To illustrate this issue, we use an example from (Zhang et al., 2023b) on a multi-modal regression problem using a two-layer fully connected network with one layer for encoding, one layer for a multi-modal head, along with concatenation fusion. A toy implementation of this example is given in Figure 1, and the corresponding implementation details are given in Appendix B.

Concretely, consider a two modality dataset  $\mathcal{D}_{mm} := \{x_i^{(m_1)}, x_i^{(m_2)}, y_i\}_{i=1}^N$  with  $x_i^{(m_1)} \in \mathbb{R}^{d_1}$ ,  $x_i^{(m_2)} \in \mathbb{R}^{d_2}$ , and  $y_i \in \mathbb{R}$  for all  $i \in [N]$ . Let the empirical input and input-output correlation matrices for modality  $m_1$  be  $C_{m_1}$  and  $C_{ym_1}$  (similarly for modality  $m_2$ ). Also let the cross-correlation matrices between  $m_1$  and  $m_2$  be  $C_{m_1 m_2}$  and  $C_{m_2 m_1}$ . The exact definitions of these matrices are given in Appendix B.

**Illustration with the two-layer model.** For the two-layer fully connected late fusion multi-modal network with concatenation fusion, we can choose parameters for uni-modal encoders for the network as  $\theta_{m_1} \in \mathbb{R}^{d_1 \times d_h}$ ,  $\theta_{m_2} \in \mathbb{R}^{d_2 \times d_h}$ , where  $d_h$  is the dimensionality of the encoder layer for each modality  $m_1$  and  $m_2$ . Note that multi-modal head  $\vartheta_{mm}$  can be partitioned to modality-specific components  $\vartheta_{mm, m_1}$  and  $\vartheta_{mm, m_2}$  as  $\vartheta_{mm} = [\vartheta_{mm, m_1}; \vartheta_{mm, m_2}]$  with  $\vartheta_{mm, m_1}, \vartheta_{mm, m_2} \in \mathbb{R}^{d_h}$ . Hence, we can denote  $\Theta_{mm} := [\theta_{m_1}; \theta_{m_2}; \vartheta_{mm, m_1}; \vartheta_{mm, m_2}]$ . All model parameters are initialized close to zero.

For a given data index  $i$ , the output of the multi-modal model can then be given as

$$\hat{y}_i = \vartheta_{mm, m_1} \theta_{m_1} x_i^{(m_1)} + \vartheta_{mm, m_2} \theta_{m_2} x_i^{(m_2)}. \quad (4)$$

Note that the decoupled nature of the output with respect to the modalities is due to the late fusion multi-modal model architecture with concatenate fusion. The corresponding regression loss can be given as  $f_{mm}(\Theta_{mm}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$ . We can then derive the gradient of  $f_{mm}$  as (see Appendix B for details)

$$\nabla_{\Theta_{mm}} f_{mm}(\Theta_{mm}) = [\theta_{m_1}^\top \Psi_1; \theta_{m_2}^\top \Psi_2; \Psi_1^\top \vartheta_{mm, m_1}; \Psi_2^\top \vartheta_{mm, m_2}], \quad (5)$$

where

$$\begin{aligned} \Psi_1 &= C_{ym_1} - \vartheta_{mm, m_1} \theta_{m_1} C_{m_1} - \vartheta_{mm, m_2} \theta_{m_2} C_{m_2 m_1}, \\ \Psi_2 &= C_{ym_2} - \vartheta_{mm, m_1} \theta_{m_1} C_{m_1 m_2} - \vartheta_{mm, m_2} \theta_{m_2} C_{m_2}. \end{aligned}$$

To ensure  $\nabla_{\Theta_{mm}} f_{mm}(\Theta_{mm}) = 0$  (i.e. stationarity), it suffices to achieve some combination of  $\Psi_1 = 0$ ,  $\Psi_2 = 0$ ,

$\theta_{m_k} = 0$  and  $\vartheta_{m_k} = 0$ . However, in general, for the model to have ‘learned’ a modality  $m_k$ , the weights corresponding to that modality should be non-zero, i.e.  $\theta_{m_k} \neq 0$  and  $\vartheta_{m_k} \neq 0$ . Thus, to achieve a stationary point that corresponds to learning both modalities, ideally, we want model parameters that satisfy  $\Psi_k = 0$ ,  $\theta_{m_k} \neq 0$  and  $\vartheta_{m_k} \neq 0$  for all  $k \in \{1, 2\}$ .

**Superficial modality preference.** Note that since the model weights are initialized from zero, the model will first visit a uni-modal stationary point where  $\Psi_k = 0$ ,  $\vartheta_{mm, m_{3-k}} = 0$  and  $\Theta_{m_{3-k}} = 0$  for some  $k \in \{1, 2\}$ , before eventually reaching the ideal stationary point that achieve  $\Psi_1 = \Psi_2 = 0$ ,  $\vartheta_{mm, m_{3-k}} \neq 0$  and  $\Theta_{m_{3-k}} \neq 0$  for all  $k \in \{1, 2\}$ . Which stationary point the model visits first will depend on the dataset statistics (Zhang et al., 2023b) (see Appendix B for more details). Furthermore, it can be shown that uni-modal stationary point the model visits first is decoupled from which modality will contribute more in minimizing  $f_{mm}$ , which is known as “superficial modality preference” (Zhang et al., 2023b). For example, as in the toy MML task depicted in Figure 1, the drop in objective value  $f_{mm}$  is smaller when modality  $m_1$  is learned (when norm of encoder weights  $\|\theta_{m_1}\|$  attain a non zero value), compared to that when modality  $m_2$  is learned, although modality  $m_1$  is learned first. Thus, imbalanced modality learning can lead to models that are overfitted to a fast learning modality, which are sub-optimal to the overall multi-modal objective.

## 2.2 Lexicographic MOO

In this section, we introduce the key MOO problem formulation that we use to alleviate the imbalance issue in MML discussed in Section 2.1. Specifically, we introduce the lexicographic MOO method that is used when one can assign apriori an order of learning objectives, based on some preference. In other words, optimizing objectives with lower preference is constrained upon the optimality of the objectives with higher preference. More concretely, consider a set of objectives  $f_m : \mathbb{R}^d \mapsto \mathbb{R}$  for  $m \in [M]$ . Then, the lexicographic MOO problem can be formulated as (Miettinen, 1999)

$$\text{lex min}_{\Theta \in \mathbb{R}^d} F^{\text{Lex}}(\Theta) := f_1(\Theta), f_2(\Theta), \dots, f_M(\Theta) \quad (6)$$

where the index of the objectives gives the order in which the optimality of each objective should be achieved. For the bi-objective case, (6) can be given as a constrained optimization problem

$$\min_{\Theta \in \mathbb{R}^d} f_2(\Theta) \quad \text{s.t.} \quad \Theta \in \arg \min_{\Theta \in \mathbb{R}^d} f_1(\Theta). \quad (7)$$

Note that this method allows one to incorporate prior knowledge about the problem into the optimization process. It can be shown that the solution of (6) is Pareto optimal (Miettinen, 1999).

## 3 Balanced MML via MOO

In this section, we first present our proposed reformulation of MML problem for addressing the imbalance issue discussed in the previous section. We then detail the corresponding algorithm development and provide a convergence analysis for the proposed method.

### 3.1 Problem formulation

In this section, we modify the original MML problem (3) to ensure balanced learning among modalities. Intuitively, the slow-to-learn modality will have the worst optimality gap. Thus, to alleviate the imbalanced learning problem, we modify the MML formulation to encourage the learning of the modality with the worst optimality gap. Specifically, we propose to achieve the optimality of multi-modal objective constrained upon the optimality of the worst-performing (in terms of optimality gap) uni-modal objective, that is

$$\begin{aligned} \min_{\Theta_{mm} \in \mathbb{R}^{d_{mm}}} f_{mm}(\Theta_{mm}) \\ \text{s.t.} \quad \{\Theta_{m_k}\} \in \arg \min_{\{\Theta_{m_k}\}} \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*), \end{aligned} \quad (8)$$

where  $\Theta_{m_k} = [\vartheta_{m_k}; \theta_{m_k}]$  with  $\vartheta_{m_k}$  being the uni-modal head dedicated to modality  $m_k$ , and  $f_{m_k}^* = \min_{\Theta_{m_k}} f_{m_k}(\Theta_{m_k})$ , for all  $k \in [K]$ .

**Remark 1.** Note that (8) follows the lexicographic MOO structure in (7), where  $f_2(\hat{\Theta}_{mm}) = f_{mm}(\Theta_{mm})$  and  $f_1(\hat{\Theta}_{mm}) = \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$  with  $\hat{\Theta}_{mm} := [\vartheta_{mm}; \vartheta_{m_1}; \dots; \vartheta_{m_K}; \theta_{m_1}; \dots; \theta_{m_K}]$ . However, unlike in (7), in (8) only part of  $\hat{\Theta}_{mm}$  (the set of uni-modal encoders) is shared between the two objectives, allowing for independent optimization of the non-shared part of  $\hat{\Theta}_{mm}$  (multi- and uni-modal heads).

Note that the shared parameters between uni-modal and multi-modal objectives are only the uni-modal encoders  $\theta_{m_k}$  for all  $k \in [K]$ . The problem (8) can be rewritten for the two modality case ( $K = 2$ ) as

$$\begin{aligned} \min_{\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}} f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}) \\ \text{s.t.} \quad \theta_{m_1}, \theta_{m_2}, \vartheta_{m_1}, \vartheta_{m_2} \\ \in \arg \min_{\theta_{m_1}, \theta_{m_2}} \max_{k \in \{1, 2\}} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*). \end{aligned} \quad (9)$$

For simplicity, in the sequel we will consider the two modality case for elaborating our proposed method, although our method can be applied for an arbitrary number of modalities ( $K > 2$ ). Note that the lexico optimization of the worst performing uni-modal objective and multi-modal objective with respect to the shared parameters  $\theta_{m_1}, \theta_{m_2}$  can be viewed as a simple bi-level optimization problem. Recently, (Shen and Chen, 2023)

introduced a reformulation of the bi-level optimization problem as a single-level one by penalizing lower-level constraints to the upper level. Leveraging this view, we can rewrite (9) as

$$\min_{\hat{\Theta}_{mm}} \hat{f}_{mm}(\hat{\Theta}_{mm}) := f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}) \quad (10)$$

$$+ \lambda \max_{k \in \{1,2\}} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*),$$

where  $\hat{\Theta}_{mm} = [\vartheta_{mm}; \vartheta_{m_1}; \vartheta_{m_2}; \theta_{m_1}; \theta_{m_2}]$ , and  $\lambda > 0$  is a problem dependent penalty parameter.

**How MIMO alleviate modality imbalance?** To understand how this new formulation alleviates modality imbalance, let us revisit the two-layer fully connected late fusion multi-modal network with concatenation fusion, now with additional uni-modal heads (linear layers) to obtain uni-modal objectives  $f_{m_1}$  and  $f_{m_2}$  (See Figure 1). Let the weight vectors corresponding to the uni-modal heads be  $\vartheta_{m_1} \in \mathbb{R}^{d_h}$  and  $\vartheta_{m_2} \in \mathbb{R}^{d_h}$ . Then, the new parameter to be optimized can be given as  $\hat{\Theta}_{mm} := [\vartheta_{mm,m_1}; \vartheta_{mm,m_2}; \vartheta_{m_1}; \vartheta_{m_2}; \theta_{m_1}; \theta_{m_2}]$ . The corresponding uni-modal objectives can be given by  $f_{m_k}(\Theta_{m_k}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i^{(m_k)})^2$  for  $k \in [2]$ , where  $\hat{y}_i^{(m_k)} = \vartheta_{m_k} \theta_{m_k} x_i^{(m_k)}$  for each data index  $i \in [N]$ . With these new definitions, we can rewrite the gradients of each layer of the multi-modal network modified with additional uni-modal heads as

$$\nabla_{\hat{\Theta}_{mm}} f_{mm}(\hat{\Theta}_{mm}) = [\theta_{m_1}^\top (\Psi_1 + \lambda_1 \check{\Psi}_1); \vartheta_{mm,m_2}^\top (\Psi_2 + \lambda_2 \check{\Psi}_2); \Psi_1^\top \vartheta_{mm,m_1}; \Psi_2^\top \vartheta_{mm,m_2}; \lambda_1 \check{\Psi}_1^\top \theta_{m_1}; \lambda_2 \check{\Psi}_2^\top \theta_{m_2}], \quad (11)$$

where  $\check{\Psi}_1 = C_{ym_1} - \vartheta_{m_1} \theta_{m_1} C_{m_1}$ ,  $\check{\Psi}_2 = C_{ym_2} - \vartheta_{m_2} \theta_{m_2} C_{m_2}$ , and  $\lambda_i = \lambda$  if  $i = \arg\max_{k \in \{1,2\}} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$ , 0 otherwise. Let us now see intuitively how this modification, introduced by incorporating modality-specific objectives, helps in balanced MML. Assume modality  $m_1$  is quick to learn. This means that initially, weights in  $m_2$  component of the model are close to zero, hence  $f_{m_2}$  is larger compared to  $f_{m_1}$ . Thus,  $\lambda_1 = 0$  and  $\lambda_2 = \lambda$ . Now, due to the amplified gradient from  $\lambda$  weighting, weights in  $m_2$  component of the model are updated rapidly via gradient contribution from the  $f_{m_2}$  objective, while modality  $m_1$  is learned via gradient contributed by  $f_{mm}$ . At stationarity, since all gradient components in (11) should be zero, gradient components contributed from  $f_{mm}$  and  $f_{m_2}$  should be each zero. Hence, the model achieves optimality for the original multi-modal objective  $f_{mm}$ , while simultaneously learning modality  $m_2$ .

### 3.2 Algorithm development

In this section, we provide the algorithm to solve the problem given in (10). First, note that the penalty component of (10) is non-smooth due to the max operator. This kind of non-smoothness results in slow convergence of  $\mathcal{O}(1/\epsilon^2)$  in subgradient-based optimization

---

#### Algorithm 1 MML-via-MOO (MIMO)

---

**input**  $\hat{\Theta}_{mm,1} := [\vartheta_{mm,1}; \vartheta_{m_1,1}; \vartheta_{m_2,1}; \theta_{m_1,1}; \theta_{m_2,1}]$ , learning rates  $\{\eta_t\}_{t=1}^T$ , penalty parameter  $\lambda$ , smoothing parameter  $\mu$   
**for**  $t = 1, \dots, T$  **do**  
     Compute gradient of  $\hat{f}_{mm}$  given in (13)  
     Update  $\hat{\Theta}_{mm,t+1}$  following (14)  
**end for**  
**output**  $\hat{\Theta}_{mm,T+1}$

---

in general, compared to  $\mathcal{O}(1/\epsilon)$  of smooth gradient-based optimization (Nesterov, 2005). To alleviate this problem, prior work has proposed to use the so-called smoothing function to incorporate prior knowledge on the non-smooth function, which has been shown to result in convergence to  $\mathcal{O}(\epsilon)$  suboptimal point of the original non-smooth problem with  $\mathcal{O}(1/\epsilon)$  iterations. We will formally define the concepts of smoothness and smoothing functions next.

**Definition 1** (Smoothness and smoothing function (Lin et al., 2024)). *A differentiable function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth if for all  $\Theta_1, \Theta_2 \in \mathbb{R}^d$ , the gradient of  $g$  satisfies the condition*

$$\|\nabla g(\Theta_1) - \nabla g(\Theta_2)\| \leq L \|\Theta_1 - \Theta_2\|. \quad (12)$$

For a continuous function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ , we call  $g_\mu : \mathbb{R}^d \rightarrow \mathbb{R}$  a smoothing function of  $g$  if for any  $\mu > 0$ ,  $g_\mu$  is continuously differentiable in  $\mathbb{R}^d$  and satisfies the conditions (1)  $\lim_{\Phi \rightarrow \Theta, \mu \downarrow 0} g_\mu(\Phi) = g(\Theta)$ ; and (2) there exists constants  $L$  and  $\alpha > 0$  independent of  $\mu$ , such that  $g_\mu$  is  $(L + \alpha\mu^{-1})$ -smooth.

Note that  $g_\mu(\Theta) := \mu \log \left( \sum_{m=1}^M \exp(\mu^{-1} g_m(\Theta)) \right)$  is a smoothing function for  $g(\Theta) = \max_{m \in [M]} g_m(\Theta)$  (Lin et al., 2024), where  $\mu > 0$  is the smoothing parameter that controls the smoothness of  $g_\mu$ . Then, we have the smoothed version of (10) as

$$\min_{\hat{\Theta}_{mm}} \hat{f}_{mm}(\hat{\Theta}_{mm}) := f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}) \quad (13)$$

$$+ \lambda \mu \log \left( \sum_{k=1}^2 \exp(\mu^{-1} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*)) \right).$$

With this formulation, we can apply gradient descent on the objective  $\hat{f}_{mm}$  given in (13), as

$$\hat{\Theta}_{mm,t+1} = \hat{\Theta}_{mm,t} - \eta_t \nabla \hat{f}_{mm}(\hat{\Theta}_{mm,t}), \quad (14)$$

where  $t$  is the iteration index, and  $\eta_t$  is the learning rate. The gradient update in 14 can be decomposed

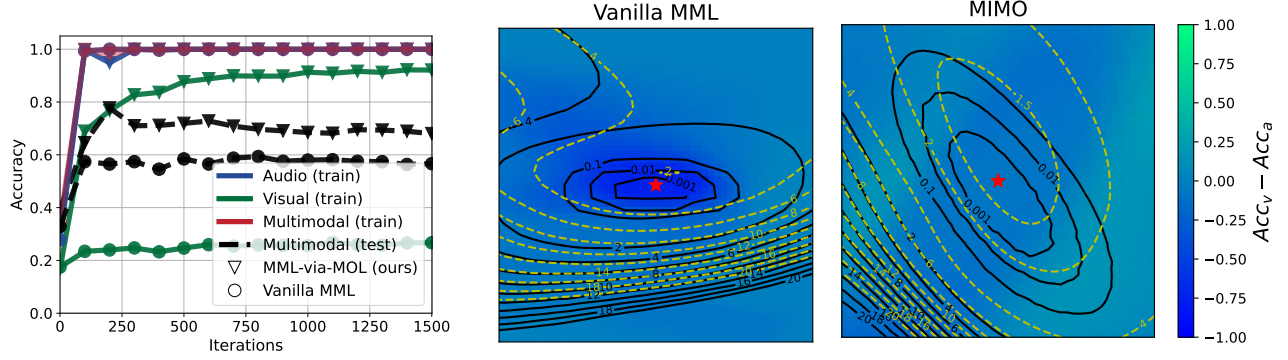


Figure 2: **Left:** Comparison of the training and testing performance of MIMO algorithm with vanilla MML (joint training with sum fusion) on CREMA-D dataset. **Middle and Right:** Comparison of the loss landscape of vanilla MML and MIMO after 1500 iterations on CREMA-D dataset. The black contours (—) denote the multi-modal training loss, and the yellow dashed contours (---) denote the multi-modal testing loss. The red star (★) denotes the convergent point of each method. The color of the heatmap denotes the difference between uni-modal training accuracies at the given point of the loss landscape, where blue (■) denotes audio modality is dominating, green (■) denotes visual modality is dominating, and higher color intensity denotes larger differences in accuracy. As illustrated by the training curves and loss landscapes, MIMO achieves lower multi-modal test loss (i.e. better generalization) by balancing the learning of each modality.

into updates of each components of  $\hat{\Theta}_{mm}$  as

$$\begin{aligned} \vartheta_{mm,t+1} &= \vartheta_{mm,t} - \eta_t \nabla_{\vartheta_{mm}} f_{mm}(\vartheta_{mm,t}, \theta_{m_1,t}, \theta_{m_2,t}) \\ \vartheta_{m_k,t+1} &= \vartheta_{m_k,t} - \eta_t \lambda \sigma_{m_k} \nabla_{\vartheta_{m_k}} f_{m_k}(\vartheta_{m_k,t}, \theta_{m_k,t}) \\ \theta_{m_k,t+1} &= \theta_{m_k,t} - \eta_t \nabla_{\theta_{m_k}} f_{mm}(\vartheta_{mm,t}, \theta_{m_1,t}, \theta_{m_2,t}) \\ &\quad - \eta_t \lambda \sigma_{m_k} \nabla_{\theta_{m_k}} f_{m_k}(\vartheta_{m_k,t}, \theta_{m_k,t}) \end{aligned}$$

where  $\sigma_{m_k} := \exp(h_{\mu,k}) / \sum_{k'=1}^2 \exp(h_{\mu,k'})$ ,  $h_{\mu,k} = \mu^{-1}(f_{m_k}(\vartheta_{m_k}, \theta_{m_k}))$ , and  $k \in \{1, 2\}$ . The MIMO algorithm is summarized in Algorithm 1. Note that the careful build-up of a single objective that accounts for balanced MML allows us to employ gradient descent on the single objective  $\hat{f}_{mm}$ , without computationally intensive heuristic subroutines that lack provable performance guarantees. Next, we provide the convergence guarantee for the proposed algorithm.

### 3.3 Theoretical analysis

In this section, we establish the convergence of the proposed methods under mild assumptions.

**Assumption 1** (Smoothness of objectives). *Functions  $f_{mm}(\Theta_{mm})$  and  $f_{m_k}(\Theta_{m_k})$  are  $L_{mm}$  and  $L_{m_k}$  smooth (Definition 1), respectively, where  $k \in [2]$ .*

**Assumption 2** (Lipschitz continuity of objectives). *For all  $k \in [2]$ , objectives  $f_{m_k}$  are  $L_{m_k,1}$  Lipschitz continuous, i.e. for any  $\Theta_{m_k}, \Theta'_{m_k}$ ,*

$$|f_{m_k}(\Theta_{m_k}) - f_{m_k}(\Theta'_{m_k})| \leq L_{m_k,1} \|\Theta_{m_k} - \Theta'_{m_k}\|. \quad (15)$$

Assumption 1 is a standard assumption on objectives used in optimization literature (Nesterov, 2018), and Assumption 2 is required in our analysis to establish the smoothness of the composite objective  $\hat{f}_{mm}$ . We can then have the following proposition.

**Proposition 1** (Smoothness of  $\hat{f}_{mm}$ ). *Under Assumptions 1 and 2, there exist  $\hat{L}_{mm} > 0$  such that  $\hat{f}_{mm}$  defined in (13) is  $\hat{L}_{mm}$ -smooth (Definition 1), where  $\hat{L}_{mm} := L_{mm} + \lambda \sum_{k=1}^2 (L_{m_k} + \mu^{-1} L_{m_k,1}^2)$ .*

Proof of Proposition 1 is given in Appendix C. Proposition 1 establishes the smoothness of the objective used in Algorithm 1 for gradient descent. From the gradient descent theory (Nesterov, 2018), we have the following.

**Theorem 1** (Convergence). *Let Assumptions 1 and 2 hold. For any  $\lambda, \mu > 0$ , and  $0 < \eta_t \leq 1/\hat{L}_{mm}$  for all  $t \in [T]$ , Algorithm 1 reaches to an  $\epsilon$  stationary point of  $\hat{f}_{mm}$  with iteration complexity of  $\mathcal{O}(1/\epsilon)$ .*

**Remark 2.** While Theorem 1 establishes the convergence of Algorithm 1 to a stationary point of the problem (13), it does not give any insight into the choice of parameters  $\lambda$  and  $\mu$  such that the optimality of the multi-modal objective  $f_{mm}$  and worst performing uni-modal objective is achieved. We provide an in-depth discussion on this in Appendix D, under additional assumptions on the problem setup.

## 4 Experiments

In this section, we empirically validate the performance of MIMO on several MML benchmarks and compare MIMO with several popular MML and MOO baselines. In addition to the performance comparison, we provide experiments to show better generalization ability of MIMO and an ablation study on MIMO parameters.

**Experiment settings.** We adopt the same experiment settings used in (Li et al., 2023) and (Peng et al., 2022) to run all the experiments on popular MML

Table 1: Multi-modal and uni-modal test accuracy performance (Acc, Acc<sub>a</sub>, Acc<sub>v</sub>, Acc<sub>t</sub>) of different MML and MOO methods on the CREMA-D and UR-Funny datasets.  $t(s)$  denotes the average subroutine time for each method. The best (highest) accuracy results are shown in **bold**. The best (lowest) subroutine time among the first three best-performing methods (in Acc) is underlined. All error values denote one standard deviation.

Method	CREMA-D				UR-Funny				
	Acc (%)	Acc <sub>a</sub> (%)	Acc <sub>v</sub> (%)	$t(s)$	Acc (%)	Acc <sub>a</sub> (%)	Acc <sub>v</sub> (%)	Acc <sub>t</sub>	$t(s)$
<b>Audio</b>	-	59.31± 0.76	-	0.028± 0.005	-	<b>58.00± 0.74</b>	-	-	0.037± 0.002
<b>Visual</b>	-	-	<b>61.04± 0.87</b>	0.029± 0.005	-	-	<b>53.12± 0.52</b>	-	0.038± 0.002
<b>Text</b>	-	-	-	-	-	-	-	<b>63.74± 1.59</b>	0.038± 0.002
<b>MML</b>	60.26± 0.84	58.02± 0.58	22.69± 1.72	0.038± 0.009	63.10± 0.59	53.02± 1.16	50.18± 0.78	62.49± 0.90	0.038± 0.002
<b>MSES</b>	57.96± 0.42	55.84± 0.91	27.37± 1.12	0.042± 0.007	62.90± 0.68	53.33± 1.17	49.91± 1.46	62.71± 0.60	0.071± 0.006
<b>MSLR</b>	62.09± 0.15	58.35± 0.62	25.62± 1.43	0.0412± 0.006	63.16± 0.45	54.77± 1.51	50.69± 0.18	61.89± 0.99	0.074± 0.012
<b>OGM-GE</b>	74.49± 0.65	53.78± 1.21	47.82± 1.51	0.112± 0.009	-	-	-	-	-
<b>AGM</b>	46.63± 0.93	43.05± 0.99	18.32± 1.11	0.205± 0.005	64.18± 0.77	54.76± 0.65	49.45± 0.90	62.74± 0.76	0.384± 0.001
<b>EW</b>	65.50± 0.50	58.80± 0.77	59.66± 1.56	0.036± 0.006	63.63± 0.42	54.00± 0.94	49.69± 0.95	62.41± 0.55	0.090± 0.002
<b>MGDA</b>	63.47± 0.79	<b>60.80± 0.68</b>	26.25± 1.19	0.310± 0.053	63.81± 0.53	53.85± 1.37	49.81± 0.80	62.96± 1.00	0.441± 0.003
<b>MMPareto</b>	68.67± 0.97	60.59± 0.57	58.82± 1.61	0.309± 0.053	63.94± 0.53	52.48± 1.57	50.31± 0.76	63.35± 1.27	0.436± 0.006
<b>MIMO</b>	<b>75.96± 0.83</b>	55.60± 1.54	59.76± 1.40	<u>0.037± 0.012</u>	<b>64.54± 0.86</b>	52.19± 1.08	50.38± 0.48	62.10 ± 1.31	<u>0.101± 0.013</u>

datasets. **CREMA-D** dataset (Cao et al., 2014) is an audio-visual dataset for speech emotion recognition, with six emotion labels. The **UR-Funny** dataset (Hasan et al., 2019) is created for humor detection, involving words (text), gestures (vision), and intonational cues (acoustic) modalities. **Kinetics-Sound** (Arandjelovic and Zisserman, 2017) is a dataset comprising 31 human action classes derived from the Kinetics dataset (Kay et al., 2017), which includes 400 categories of YouTube videos featuring both audio and visual components. **VGGSound** (Chen et al., 2020) is a large-scale video dataset with 309 classes, representing a broad spectrum of everyday audio events. See additional experiment details and results for **AV-MNIST**, **AVE**, and **CMU-MOSEI** datasets given in Appendix F. We compare MIMO with several popular MML baselines such as Modality-Specific Early Stopping (**MSES**) (Fujimori et al., 2020), Modality-Specific Learning Rate (**MSLR**) (Yao and Mihalcea, 2022), On-the-fly Gradient Modulation Generalization Enhancement (**OGM-GE**) (Peng et al., 2022) (this method is designed only for the two modality case), and Adaptive Gradient Modulation (**AGM**) (Li et al., 2023) methods. In addition to these baselines, we also compare with MTL baselines like equal weighting (**EW**), Multiple gradient descent algorithm (**MGDA**) (Désidéri, 2012), and **MMPareto** (Wei and Hu, 2024) for solving the MML problem as MOO.

**Balanced MML for better generalization.** First, we provide qualitative results on how MIMO improves the generalization performance in real-world MML tasks, using CREMA-D dataset. Figure 2 Left shows the learning behavior for vanilla MML and MIMO. It can be seen that while MIMO learns the slow-to-learn visual modality, vanilla MML overfits the audio modality which results in poor testing performance.

Furthermore, in Figure 2 Middle and Right we investigate the loss landscape around the model trained using vanilla MML and the proposed MIMO methods, in reduced dimensionality (Li et al., 2018). It can be seen that vanilla MML, although having lower training losses (solid black contours) compared to MIMO, has poor testing loss (dashed yellow contours). Furthermore, it can be seen that the training accuracy disparity between audio and visual modalities (blue-green shade) is very high in favor of audio modality. On the other hand, MIMO has a more balanced training accuracy performance between audio and visual modalities, and has a better testing loss performance compared to vanilla MML, at the expense of slightly poor performance in training loss. This suggests that balanced MML prevents the model from overfitting to a specific modality, thereby improving its generalization.

**Comparison with baselines.** Next, we demonstrate the performance gain from the MIMO method in real-world MML benchmarks over existing balanced MML methods and MOO methods. Table 1 shows the performance of MIMO compared to existing MML and MOO baselines in the MML classification benchmarks CREMA-D and UR-Funny. We first compare the performance in the CREMA-D benchmark. It can be seen that MIMO achieves the best test accuracy performance compared to the baselines. By comparing the results obtained for individual modality training, it can be seen that MIMO can achieve superior performance via balanced MML learning, whereas vanilla MML fails to even perform comparably to the best-performing individual modality (visual modality). Furthermore, it can be seen that naively applying MOO methods on multi-modal and uni-modal objectives (e.g., EW and MGDA) does not improve the multi-modal performance, as there is no fine-grained control between



Table 2: Comparison using VGGSound dataset.

	Acc (%)	Acc <sub>a</sub> (%)	Acc <sub>v</sub> (%)	$t(s)$
MML	60.8 $\pm$ 0.13	42.83 $\pm$ 1.04	15.43 $\pm$ 1.18	0.011 $\pm$ 0.001
OGM-GE	62.13 $\pm$ 1.31	32.50 $\pm$ 2.26	22.00 $\pm$ 0.01	0.121 $\pm$ 0.007
EW	63.90 $\pm$ 1.58	<b>48.77</b> $\pm$ 2.01	25.20 $\pm$ 1.40	<u>0.006</u> $\pm$ 0.001
MMPareto	66.07 $\pm$ 1.04	48.07 $\pm$ 1.37	28.87 $\pm$ 1.44	0.389 $\pm$ 0.053
MIMO	<b>69.10</b> $\pm$ 1.13	41.47 $\pm$ 1.04	<b>38.20</b> $\pm$ 1.01	<u>0.019</u> $\pm$ 0.004

Table 3: Comparison using Kinetics-Sound dataset.

	Acc (%)	Acc <sub>a</sub> (%)	Acc <sub>v</sub> (%)	$t(s)$
MML	59.83 $\pm$ 1.78	41.2 $\pm$ 6.34	18.67 $\pm$ 1.56	0.023 $\pm$ 0.007
OGM-GE	63.73 $\pm$ 1.37	44.10 $\pm$ 0.01	22.57 $\pm$ 2.56	0.247 $\pm$ 0.119
EW	60.73 $\pm$ 1.77	45.13 $\pm$ 2.56	33.3 $\pm$ 2.34	<u>0.026</u> $\pm$ 0.010
MMPareto	68.60 $\pm$ 1.41	<b>48.07</b> $\pm$ 1.37	37.23 $\pm$ 1.37	0.689 $\pm$ 0.098
MIMO	<b>69.60</b> $\pm$ 1.41	45.07 $\pm$ 1.04	<b>43.13</b> $\pm$ 1.78	<u>0.039</u> $\pm$ 0.015

uni-modal/multi-modal objective optimization. For example, it can be seen that MGDA is heavily biased towards the audio modality, leading to poor performance in multi-modal accuracy.

We then compare the performance of MML and MOO baselines in the three-modality benchmark dataset UR-Funny and observe that MIMO performs comparably or better compared the baselines, with a subroutine time close to vanilla MML, which is consistent with the observations in the CREMA-D dataset. Next, we compare the performance using the Kinetics-Sound benchmark. The results are given in Table 3. It can be seen that MIMO outperforms MOO and balanced MML baselines. We attribute this superior performance to balanced learning in different modalities, as evidenced from the lower disparity in uni-modal accuracies for audio and visual modalities (only  $\sim 3\%$  difference for MIMO, while the next smallest disparity  $\sim 11\%$  is from MMPareto). Similar observations can be made from the experiment results for VGGSound benchmark, given in Table 2.

Furthermore, in most of the above the experiments, MIMO has the fastest subroutine times (underlined values in the  $t(s)$  column) in the top three best performing methods, and is in the same order as vanilla MML. MIMO achieves a best of  $\sim \times 20$  speed-up compared to the next best performing baseline (MMPareto) in VGGSound benchmark experiments. While MIMO consistently outperforms baselines in all datasets, it can be seen that, for some datasets like UR-FUNNY, simple methods like vanilla MML also work reasonably well. This can be due to the nature of the dataset, which does not satisfy conditions for imbalance in MML (Lu, 2024), and hence does not have a severe imbalance in learning modalities. However, since MIMO does not have a significant computational overhead, we believe applying MIMO in such cases is not undesirable.

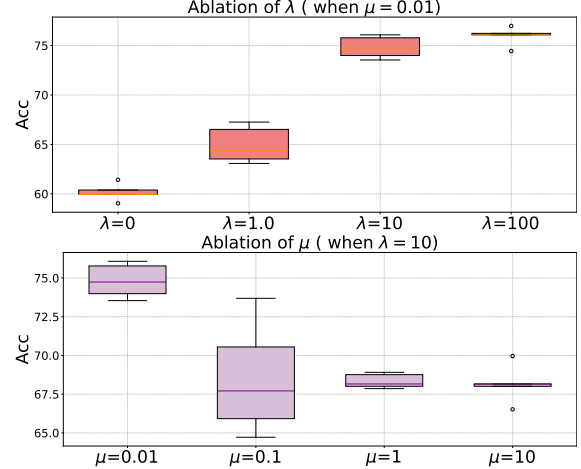


Figure 3: Ablation of hyperparameters.

**Ablation of MIMO parameters** In Figure 3, we provide an ablation for the choice of  $\lambda$  and  $\mu$  for MIMO using CREMA-D dataset. It can be seen that, for very small  $\lambda$  ( $= 1.0$ ), the performance is poor compared to larger  $\lambda$  ( $\geq 10$ ), which is expected, since the impact of constraint is weaker for small  $\lambda$ . We also see that increasing  $\lambda$  can improve performance, but the performance increase is marginal for larger  $\lambda$  ( $\geq 10$ ). Considering the choice of  $\mu$ , we see that for smaller  $\mu$ , MIMO performs well, whereas when  $\mu$  is larger, performance degrades significantly. This is also expected, as when  $\mu$  is large, smoothed max deviates too much from the max function, which will prevent MIMO from prioritizing the worst performing uni-modal objective.

## 5 Conclusions and Future Work

In this paper, we proposed a new problem formulation for balanced MML that prefers learning the worst-performing uni-modal objective over the multi-modal objective. We motivated the proposed method and why the proposed method can alleviate the imbalanced learning issue in MML while optimizing the multi-modal objective. We then proposed MIMO, a simple gradient-based algorithm to solve the modified problem which does not involve computationally intensive subroutines and has convergence guarantees. Empirical evaluation of the proposed method in MML benchmarks shows that it can outperform existing balanced MML methods, validating the efficacy of the proposed method. This work only focuses on late fusion cases with concatenation/sum fusion, and hence does not verify the efficacy of MIMO in more complex multi-modal model architectures. As future work, it would be interesting to see how MIMO-like algorithms that incorporate MOO tools to alleviate modality imbalance can be designed for early or hybrid fusion methods.



## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.
- V Joseph Bowman Jr. On the relationship of the tchebycheff norm and the efficient frontier of multiple-criteria objectives. In *Multiple Criteria Decision Making: Proceedings of a Conference Jouy-en-Josas, France May 21–23, 1975*, pages 76–86. Springer, 1976.
- Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.
- Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.
- Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *arXiv preprint arXiv:2305.20057*, 2023.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proc. of International Conference on Machine Learning*, virtual, July 2018.
- Corinna Cortes, Mehryar Mohri, Javier Gonzalvo, and Dmitry Storcheus. Agnostic learning with multiple objectives. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pages 20485–20495, virtual, 2020.
- Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pages 8632–8656. PMLR, 2023.
- Jean-Antoine Désidéri. Multiple-gradient Descent Algorithm (MGDA) for Multi-objective Optimization. *Comptes Rendus Mathématique*, 350(5-6), 2012.
- Heshan Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent stochastic approach. In *Proc. of International Conference on Learning Representations*, Kigali, Rwanda, 2023.
- Jörg Fliege, A Ismael F Vaz, and Luís Nunes Vicente. Complexity of Gradient Descent for Multi-objective Optimization. *Optimization Methods and Software*, 34(5):949–959, 2019.
- Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-specific learning rate control for multimodal classification. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pages 412–422. Springer, 2020.
- Yu Geng, Zongbo Han, Changqing Zhang, and Qinghua Hu. Uncertainty-aware multi-view representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7545–7553, 2021.
- Xiang Gu, Xi Yu, Jian Sun, Zongben Xu, et al. Adversarial reweighting for partial domain adaptation. In *Proc. Advances in Neural Info. Process. Syst.*, virtual, December 2021.
- Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision*, Munich, Germany, July 2018.
- Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566, 2022.
- Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftexhar Tanveer, Louis-Philippe Morency, et al. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*, 2019.
- Pengbo Hu, Xingyu Li, and Yi Zhou. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*, 2022.
- Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can

- achieve modality reconciliation. *arXiv preprint arXiv:2405.09321*, 2024.
- Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022.
- Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. *Advances in neural information processing systems*, 30, 2017.
- Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating modality collapse in multimodal vaes via impartial optimization. In *International Conference on Machine Learning*, pages 9938–9964. PMLR, 2022.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- A Kendall, Y Gal, and R Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint:1705.07115*, 2017.
- Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.
- Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023.
- Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. *arXiv preprint arXiv:2402.19078*, 2024.
- Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Proc. Advances in Neural Info. Process. Syst.*, virtual, December 2021.
- Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.
- Suyun Liu and Luis Nunes Vicente. The Stochastic Multi-gradient Algorithm for Multi-objective Optimization and its Application to Supervised Machine Learning. *Annals of Operations Research*, pages 1–30, 2021.
- Zhou Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- Mengmeng Ma, Jian Ren, Long Zhao, Davide Testugine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.
- Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.
- Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.
- Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
- James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. Parts of speech-grounded subspaces in vision-language models. *Advances in Neural Information Processing Systems*, 36: 2700–2724, 2023.
- Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022.
- Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.
- Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proc. Advances in Neural Info. Process. Syst.*, Montreal, Canada, December 2018.
- Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015. PMLR, 2023.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. AlfworlD: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018.

- Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.
- Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R Selvaraju, Chetan Ramaiah, Ran Xu, Joseph F JaJa, and Larry S Davis. Tag: Boosting text-vqa via text-aware visual question-answer generation. *arXiv preprint arXiv:2208.01813*, 2022.
- Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.
- Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, et al. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. *arXiv preprint arXiv:2403.10667*, 2024.
- Yake Wei and Di Hu. Mmpareto: boosting multimodal learning with innocent unimodal assistance. In *International Conference on Machine Learning*, 2024.
- Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.
- Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.
- Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834, 2022.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Proc. Advances in Neural Info. Process. Syst.*, virtual, December 2020.
- AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pages 2236–2246, 2018.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023a.
- Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466, 2024.
- Yedi Zhang, Peter E Latham, and Andrew Saxe. A theory of unimodal bias in multimodal learning. *arXiv preprint arXiv:2312.00935*, 2023b.
- Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2019.
- Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie Gu, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In *Proc. Advances in Neural Information Processing Systems*, volume 35, pages 38103–38115, New Orleans, LA, December 2022.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

## Supplementary Material for “Mitigating Modality Imbalance in Multi-modal Learning via Multi-objective Optimization”

### A Notations

A summary of notations used in this work is listed in Table 4 for ease of reference.

Table 4: Notations and their descriptions.

Notation	Description
$K$	Number of modalities considered. We use $K=2$ in most parts of the paper for conciseness
$N$	Number of datapoints in the multi-modal dataset $\mathcal{D}_{mm}$
$k$	Index used to denote modality, $k \in [K]$
$i$	Index used to denote datapoint, $i \in [N]$
Dataset	
$x_i^{(m_k)}$	Input corresponding to modality $m_k$ for the datapoint index $i$
$y_i$	Target output for the datapoint index $i$
$\hat{y}_i$	Multi-modal output of the vanilla MML/MIMO model for the datapoint index $i$
$\hat{y}_i^{(m_k)}$	Uni-modal output of the MIMO model for the modality $m_k$ for the datapoint index $i$
Model	
$\vartheta_{mm}$	Multi-modal head in the vanilla MML/MIMO model (see Figure 1)
$\vartheta_{m_k}$	Uni-modal head for the modality $m_k$ in the MIMO model (see Figure 1)
$\theta_{m_k}$	Uni-modal encoder for the modality $m_k$ in the vanilla MML/MIMO model (see Figure 1)
$\Theta_{mm}$	Concatenation of all the components in the vanilla MML model, i.e. $\Theta_{mm} := [\vartheta_{mm}; \theta_{m_1}; \theta_{m_2}; \dots; \theta_{m_K}]$
$\hat{\Theta}_{mm}$	Concatenation of the MIMO model parameters, i.e. $\hat{\Theta}_{mm} := [\vartheta_{mm}; \vartheta_{m_1}; \vartheta_{m_2}; \dots; \vartheta_{m_K}; \theta_{m_1}; \theta_{m_2}; \dots; \theta_{m_K}]$
$\Theta_{m_k}$	Concatenation of all the $m_k$ modality specific components in the MIMO model, i.e. $\Theta_{m_k} := [\vartheta_{m_k}; \theta_{m_k}]$
Objectives	
$f_{mm}(\Theta_{mm})$	Vanilla multi-modal objective
$f_{m_k}(\Theta_{m_k})$	Uni-modal objective for modality $m_k$ induced by uni-modal head $\vartheta_{m_k}$ in MIMO model
$\hat{f}_{mm}(\hat{\Theta}_{mm})$	MIMO objective, which is a combination of $f_{mm}$ and $f_{m_k}$ for all $k \in [K]$ (defined in (13) for $K=2$ )
Toy Illustration	
$\vartheta_{mm, m_k}$	Parameter matrix for the partition of $\vartheta_{mm}$ that corresponds to the modality $m_k$
$C_{m_k}$	Empirical input correlation matrix for modality $m_k$ (defined in Section B)
$C_{m_k m_{3-k}}$	Empirical cross-correlation matrix between modality $m_k$ and modality $m_{3-k}$ (defined in Section B)
$C_{ym_k}$	Empirical input-output correlation matrix for modality $m_k$ (defined in Section B)

Let the empirical input and input-output correlation matrices for modality  $m_1$  be  $C_{m_1}$  and  $C_{ym_1}$  (similarly for modality  $m_2$ ). Also let the cross-correlation matrices between  $m_1$  and  $m_2$  be  $C_{m_1 m_2}$  and  $C_{m_2 m_1}$ .

### B Details of Toy Example

In this section, we provide the implementation details of the toy experiment used to generate the learning curves given in Figure 1. This experiment is motivated by a similar illustration given in (Zhang et al., 2023b).

**Datset.** To generate multi-modal data  $\mathcal{D}_{mm} := \{x_i^{(m_1)}, x_i^{(m_2)}, y_i\}_{i=1}^N$ , we sample each element of  $x_i^{(m_1)}$  from  $\mathcal{N}(0, 25)$  and each element of  $x_i^{(m_2)}$  from  $\mathcal{N}(0, 0.25)$ , where  $x_i^{(m_1)}, x_i^{(m_2)} \in \mathbb{R}^{50}$ ,  $\mathcal{N}(\mu, \sigma^2)$  are Gaussian distributions with mean  $\mu$  and variance  $\sigma^2$ . We set the number of data points  $N = 700$ . The label for each datapoint is generated as  $y_i = 0.001x_i^{(m_1)} + x_i^{(m_2)}$ . The dataset is generated in this way so that it satisfies the condition for superficial modality preference given by

$$\|C_{ym_1}\| > \|C_{ym_2}\| \quad \text{and} \quad C_{ym_1} C_{m_1}^{-1} C_{ym_1}^\top < C_{ym_2} C_{m_2}^{-1} C_{ym_2}^\top, \quad (16)$$

Table 5: Gradient of the objective for each layer of the network for vanilla MML (column  $f_{mm}(\Theta)$ ) and MIMO (column  $f_{mm}(\Theta)$  + column  $\Delta$ ). Column  $\Delta$  contain the additional gradient components for  $\hat{f}_{mm}(\hat{\Theta}_{mm})$  (if any) compared to that of  $f_{mm}(\Theta)$ . Rows  $\nabla_{\vartheta_{m_1}}$  and  $\nabla_{\vartheta_{m_2}}$  for column  $f_{mm}(\Theta)$  is empty because only MIMO model  $\hat{\Theta}_{mm}$  contains  $\vartheta_{m_1}$  and  $\vartheta_{m_2}$ . In column  $\Delta$ ,  $\lambda_i = \lambda$  if  $i = \arg \max_{i \in \{1,2\}} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*)$ , else 0, where  $\lambda$  is the penalty parameter in (9).

	$f_{mm}(\Theta)$	$\Delta = \hat{f}_{mm}(\hat{\Theta}_{mm}) - f_{mm}(\Theta)$
$\nabla_{\theta_{m_1}}$	$\vartheta_{mm,m_1}^\top (C_{ym_1} - \vartheta_{mm,m_1} \theta_{m_1} C_{m_1} - \vartheta_{mm,m_2} \theta_{m_2} C_{m_2 m_1})$	$\lambda_1 \vartheta_{m_1}^\top (C_{ym_1} - \vartheta_{m_1} \theta_{m_1} C_{m_1})$
$\nabla_{\theta_{m_2}}$	$\vartheta_{mm,m_2}^\top (C_{ym_2} - \vartheta_{mm,m_1} \theta_{m_1} C_{m_1 m_2} - \vartheta_{mm,m_2} \theta_{m_2} C_{m_2})$	$\lambda_2 \vartheta_{m_2}^\top (C_{ym_2} - \vartheta_{m_2} \theta_{m_2} C_{m_2})$
$\nabla_{\vartheta_{mm,m_1}}$	$(C_{ym_1} - \vartheta_{mm,m_1} \theta_{m_1} C_{m_1} - \vartheta_{mm,m_2} \theta_{m_2} C_{m_2 m_1}) \theta_{m_1}^\top$	—
$\nabla_{\vartheta_{mm,m_2}}$	$(C_{ym_2} - \vartheta_{mm,m_1} \theta_{m_1} C_{m_1 m_2} - \vartheta_{mm,m_2} \theta_{m_2} C_{m_2}) \theta_{m_2}^\top$	—
$\nabla_{\vartheta_{m_1}}$	—	$\lambda_1 (C_{ym_1} - \vartheta_{m_1} \theta_{m_1} C_{m_1}) \theta_{m_1}^\top$
$\nabla_{\vartheta_{m_2}}$	—	$\lambda_2 (C_{ym_2} - \vartheta_{m_2} \theta_{m_2} C_{m_2}) \theta_{m_2}^\top$

where  $C_{m_k} := \frac{1}{N} \sum_{i=1}^N x_i^{(m_k)} (x_i^{(m_k)})^\top$ ,  $C_{m_k m_{3-k}} := \frac{1}{N} \sum_{i=1}^N x_i^{(m_k)} (x_i^{(m_{3-k})})^\top$ , and  $C_{ym_k} := \frac{1}{N} \sum_{i=1}^N y_i (x_i^{(m_k)})^\top$ . The derivation of the condition (16) for superficial modality preference follow the derivation steps given in (Zhang et al., 2023b) Appendix F.

**Models.** For the vanilla multi-modal model (as shown in Figure 1 (a)) we use a linear layer of input size 50 and output size 100 as the modality encoder for each modality  $m_1$  and  $m_2$ , and then use linear layers of input size 100 and output size 1 for the modality-specific multi-modal head part for each modality  $m_1$  and  $m_2$ . Finally, the multi-modal output is obtained by summing the output of each modality-specific part. For the MIMO model, we use the same architecture for the multi-modal part of the model, and use two additional linear layers, each with input size 100 and output size 1 to generate uni-modal outputs (as shown in Figure 1 (b)).

**Optimization.** We use a learning rate of 0.01 for both vanilla MML and MIMO methods. For the MIMO method, we set  $\lambda = 10$  and  $\mu = 0.2$ . The expressions of the gradients used in vanilla MML and MIMO are summarized in Table 5. The derivation of gradients follow a similar approach to that given in (Zhang et al., 2023b) Appendix A.

**Superficial modality preference.** In Figure 1 (a), we can see that vanilla MML is quick to learn modality  $m_1$ . However, it does not contribute to minimizing the multi-modal objective compared to modality  $m_2$ . This phenomenon is known as “superficial modality preference” (Zhang et al., 2023b), and occurs due to the properties of the dataset. More concretely, let the time taken to reach  $\mathcal{M}_{m_1}$  and  $\mathcal{M}_{m_2}$  are  $t_{m_1}$  and  $t_{m_2}$ , respectively. Furthermore, let the objective value at the manifolds  $\mathcal{M}_{m_1}$  and  $\mathcal{M}_{m_2}$  be  $f_{mm}(\mathcal{M}_{m_1})$  and  $f_{mm}(\mathcal{M}_{m_2})$ , respectively. Then we say model has “superficial modality preference” (Zhang et al., 2023b) if the following condition holds

$$t_{m_1} < t_{m_2} \quad \text{and} \quad f_{mm}(\mathcal{M}_{m_1}) > f_{mm}(\mathcal{M}_{m_2}). \quad (17)$$

It can be shown that if the dataset statistics satisfy the following condition, applying SGD on  $f_{mm}$  with model  $\Theta$  will result in superficial modality preference:

$$\|C_{ym_1}\| > \|C_{ym_2}\| \quad \text{and} \quad C_{ym_1} C_{m_1}^{-1} C_{ym_1}^\top < C_{ym_2} C_{m_2}^{-1} C_{ym_2}^\top. \quad (18)$$

Note that the condition depends only on the statistics of each modality data. Thus, applying SGD on multi-modal objective  $f_{mm}$  parameterized by a late fusion multi-modal  $\Theta$  can result in the model giving priority to one modality, which may not be contributing most in minimizing the objective. The data set used in this toy example is generated in such a way that the above conditions for superficial modality preference are met.

## C Proof of Proposition 1.

In this section, we provide the proof for Proposition 1.

*Proof.* Consider any  $\hat{\Theta}_{mm} = [\vartheta_{mm}; \vartheta_{m_1}; \vartheta_{m_2}; \theta_{m_1}; \theta_{m_2}]$ , with  $\Theta_{m_k} = [\vartheta_{m_k}; \theta_{m_k}]$  for  $k \in \{1, 2\}$ . For brevity, we will omit the argument of the function/gradient in derivation; for example  $\nabla_{\hat{\Theta}_{mm}} f_{mm}(\hat{\Theta}_{mm})$  will be denoted

as  $\nabla_{\hat{\Theta}_{mm}} f_{mm}$ . However, we will carefully consider the dependence of the function on the corresponding parameter, when we take gradients. Furthermore, we will denote the dimension of a vector parameter  $v$  as  $\dim(v)$ . Our goal in this proof is to show that  $\nabla_{\hat{\Theta}_{mm}}^2 \hat{f}_{mm} \preceq \hat{L}_{mm} I_0$  for some  $\hat{L}_{mm} > 0$ , where  $\nabla_{\hat{\Theta}_{mm}}^2 \hat{f}_{mm}$  is the Hessian of  $\hat{f}_{mm}$ , and  $I_0 \in \mathbb{R}^{\dim(\hat{\Theta}_{mm}) \times \dim(\hat{\Theta}_{mm})}$  is an identity matrix. We first derive the gradient of  $g_\mu := \mu \log \left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \right)$  with respect to  $\hat{\Theta}_{mm}$ . We have

$$\begin{aligned} \nabla_{\hat{\Theta}_{mm}} g_\mu &= \frac{\mu}{\sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right)} \cdot \sum_{k=1}^2 \frac{1}{\mu} \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \\ &= \frac{1}{\sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right)} \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \nabla_{\hat{\Theta}_{mm}} f_{m_k}. \end{aligned} \quad (19)$$

Then we can compute  $\nabla_{\hat{\Theta}_{mm}}^2 g_\mu$  as

$$\begin{aligned} \nabla_{\hat{\Theta}_{mm}}^2 g_\mu &= \nabla_{\hat{\Theta}_{mm}} \left( \frac{1}{\sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right)} \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right) \\ &= \frac{\left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \right) \nabla_{\hat{\Theta}_{mm}} \left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right)}{\left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \right)^2} \\ &\quad - \frac{\left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right) \nabla_{\hat{\Theta}_{mm}} \left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \right)}{\left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k} - f_{m_k}^*}{\mu} \right) \right)^2} \\ &= \sum_{k=1}^2 \Psi(z_k) \left( \frac{1}{\mu} \nabla_{\hat{\Theta}_{mm}} f_{m_k} \nabla_{\hat{\Theta}_{mm}} f_{m_k}^\top + \nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} \right) \\ &\quad - \frac{1}{\mu} \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right) \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right)^\top, \end{aligned} \quad (20)$$

where  $z_k = \frac{f_{m_k} - f_{m_k}^*}{\mu}$ , and  $\Psi$  is the softmax operator given by  $\Psi(z_i) = \frac{\exp(z_i)}{\sum_{k=1}^2 \exp(z_k)}$ . We then rewrite (20) as

$$\begin{aligned} \nabla_{\hat{\Theta}_{mm}}^2 g_\mu &= \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} + \frac{1}{\mu} \left[ \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \nabla_{\hat{\Theta}_{mm}} f_{m_k}^\top \right. \\ &\quad \left. - \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right) \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right)^\top \right]. \end{aligned} \quad (21)$$

Now, consider  $\nabla_{\hat{\Theta}_{mm}}^2 f_{m_k}$  for  $k \in \{1, 2\}$ . Since  $f_{m_k}$  is  $L_{m_k}$ -smooth (Assumption 1), we have

$$\begin{aligned} &\nabla_{\hat{\Theta}_{m_k}}^2 f_{m_k} \preceq L_{m_k} I_k, \quad I_k \in \mathbb{R}^{\dim(\Theta_{m_k})} \text{ is an identity matrix} \\ \Rightarrow &\nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} \preceq L_{m_k} I_0 \\ \Rightarrow &v^\top (\nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} - L_{m_k} I_0) v \leq 0 \quad \text{for any } v \in \mathbb{R}^{\dim(\hat{\Theta}_{mm})} \\ \Rightarrow &\sum_{k=1}^2 \Psi(z_k) v^\top (\nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} - L_{m_k} I_0) v \leq 0 \\ \Rightarrow &v^\top \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} - \sum_{k=1}^2 L_{m_k} I_0 \right) v \leq 0 \\ \Rightarrow &\sum_{k=1}^2 \Psi \nabla_{\hat{\Theta}_{mm}}^2 f_{m_k} \preceq \sum_{k=1}^2 L_{m_k} I_0. \end{aligned} \quad (22)$$

Next, considering the second term of (21), we have for any  $v \in \mathbb{R}^{\dim(\hat{\Theta}_{mm})}$ ,

$$\begin{aligned}
 & v^\top \left[ \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \nabla_{\hat{\Theta}_{mm}} f_{m_k}^\top - \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right) \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right)^\top \right. \\
 & \quad \left. - \sum_{k=1}^2 L_{m_k,1}^2 I_0 \right] v \\
 &= \sum_{k=1}^2 \Psi(z_k) y_k^2 - \left( \sum_{k=1}^2 \Psi(z_k) y_k \right)^2 - \|v\|^2 \sum_{k=1}^2 L_{m_k,1}^2, \quad y_k = v^\top \nabla_{\hat{\Theta}_{mm}} f_{m_k} \text{ for } k \in \{1, 2\} \\
 &\leq \sum_{k=1}^2 \Psi(z_k) y_k^2 - \|v\|^2 \sum_{k=1}^2 L_{m_k,1}^2 \\
 &\leq \sum_{k=1}^2 \Psi(z_k) \|\nabla_{\hat{\Theta}_{mm}} f_{m_k}\|^2 \|v\|^2 - \|v\|^2 \sum_{k=1}^2 L_{m_k,1}^2, \quad \text{due to Cauchy-Schwarz inequality} \tag{23}
 \end{aligned}$$

$$\leq \|v\|^2 \sum_{k=1}^2 \Psi(z_k) (\|\nabla_{\hat{\Theta}_{mm}} f_{m_k}\|^2 - L_{m_k,1}^2) \tag{24}$$

$$\leq 0, \tag{25}$$

where the last inequality is due to Assumption 2. The above inequality suggests that

$$\sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \nabla_{\hat{\Theta}_{mm}} f_{m_k}^\top - \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right) \left( \sum_{k=1}^2 \Psi(z_k) \nabla_{\hat{\Theta}_{mm}} f_{m_k} \right)^\top \preceq \sum_{k=1}^2 L_{m_k,1}^2 I_0. \tag{26}$$

Putting together (21), (22), and (26), we have

$$\nabla_{\hat{\Theta}_{mm}}^2 g_\mu \preceq \sum_{k=1}^2 \left( L_{m_k} + \frac{L_{m_k,1}^2}{\mu} \right) I_0. \tag{27}$$

On the other hand, from the  $L_{mm}$ -smoothness of  $f_{mm}$  (Assumption 1), we have

$$\nabla_{\hat{\Theta}_{mm}}^2 f_{mm} \preceq L_{mm} I_0. \tag{28}$$

Since  $\hat{f}_{mm} = f_{mm} + \lambda g_\mu$ , we can have

$$\nabla_{\hat{\Theta}_{mm}}^2 \hat{f}_{mm} \preceq \hat{L}_{mm} I_0, \tag{29}$$

where  $\hat{L}_{mm} := L_{mm} + \lambda \sum_{k=1}^2 \left( L_{m_k} + \frac{L_{m_k,1}^2}{\mu} \right) > 0$ , which completes the proof.  $\square$

## D Analysis of Solutions from Algorithm 1

In Section 3.3, we provided a convergence guarantee of Algorithm 1 to the stationary point of the objective (13) under standard assumptions on the underlying uni-modal and multi-modal objectives. In this section, we provide an in-depth analysis of the choice of  $\mu$  and  $\lambda$  such that the constraint of worse-performing uni-modal objective achieves some desired optimality. To this end, we follow the theoretical analysis given in (Shen and Chen, 2023) on general penalty-based gradient descent (PBGD) for the BLO problem.

First, we make the following assumption on the objectives  $f_{mm}$ ,  $f_{m_1}$ , and  $f_{m_2}$ .

**Assumption 3.**  $\vartheta_{mm}, \vartheta_{m_1}, \vartheta_{m_2}$  are fixed such that  $f_{mm}$ ,  $f_{m_1}$ , and  $f_{m_2}$  only depends on  $\theta_{m_1}$ ,  $\theta_{m_2}$ .

While Assumption 3 is restrictive towards the choice of model architectures, this kind of assumption are used in prior literature when analysing the learning behavior of deep neural networks (Huang et al., 2022).



Following Assumption 3, we consider the  $\epsilon$ -approximate smoothed version of the original constrained optimization problem 9, given by

$$\begin{aligned} \min_{\theta_{m_1}, \theta_{m_2}} \quad & f_{mm}(\theta_{m_1}, \theta_{m_2}) \\ \text{s.t.} \quad & \mu \log \left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \right) \right) - \mu \log 2 \leq \epsilon, \end{aligned} \quad (30)$$

noting that  $\min_{\theta_{m_1}, \theta_{m_2}} \mu \log \left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \right) \right) = \mu \log 2$ . Then we can write the value based penalized problem (Shen and Chen, 2023) corresponding to 30 as

$$\min_{\theta_{m_1}, \theta_{m_2}} f_{mm}(\theta_{m_1}, \theta_{m_2}) + \lambda \mu \log \left( \frac{1}{2} \sum_{k=1}^2 \exp \left( \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \right) \right). \quad (31)$$

Note that the factor of  $\frac{1}{2}$  inside the log in penalty term is a constant, and does not affect the optimizer. Thus, it is clear that Algorithm 1 optimizes the objective given in 31, if we fix the uni-modal and multi-modal heads. Next, we derive a relationship between the solution of problem 31 and problem 30, which gives an insight into choice of parameters  $\lambda$  and  $\mu$  in Algorithm 1 to satisfy the optimality of the multi-modal objective  $f_{mm}$  and the optimality of worst performing uni-modal objective.

**Assumption 4.** For all  $k \in [2]$ , there exist constants  $\mu_k$  such that the following inequality holds

$$\|\nabla f_{m_k}(\theta_{m_k})\|^2 \geq \frac{1}{\mu_k} (f_{m_k}(\theta_{m_k}) - f_{m_k}^*) \quad (32)$$

When the above condition hold, we say the function  $f_{m_k}$  satisfies the Polyak-Lojasiewicz (PL) condition with modulus  $\frac{1}{\mu_k}$ , or  $f_{m_k}$  is  $\frac{1}{\mu_k}$ -PL.

The above assumption is reasonable in the context of deep learning since it has been shown that over-parameterized neural networks can lead to losses that satisfy the PL inequality (Liu et al., 2022).

**Assumption 5.** For any  $\mu > 0$ , there exists a constant  $C_\mu > 0$ , where

$$C_\mu := \max_{\theta_1, \theta_2} \frac{\sum_{k=1}^2 \exp \left( \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \right)}{\min_{k \in [2]} \exp \left( \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \right)}. \quad (33)$$

Assumption 5 assumption can be true, if for example both objectives  $f_{m_1}$  and  $f_{m_2}$  are upperbounded. While this is not a realistic assumption in general, this assumption can generally hold for a local area of model initialization. Since we provide the optimality guarantees in Theorem 2 for local solutions, we believe this is a reasonable assumption in this context.

**Assumption 6.**  $f_{mm}$  is  $L_{1,mm}$ -Lipschitz continuous, i.e. there constant  $L_{1,mm} > 0$  such that for any  $\Theta = [\theta_{m_1}; \theta_{m_2}]$  and  $\Theta' = [\theta'_{m_1}; \theta'_{m_2}]$ , we have

$$|f_{mm}(\Theta) - f_{mm}(\Theta')| \leq L_{1,mm} \|\Theta - \Theta'\| \quad (34)$$

The above assumption is standard in non-convex optimization literature. We now show the lower level constraint function satisfies the PL condition, given the above assumptions.

**Proposition 2.** Let Assumptions 3-5 hold. If we chose  $\mu \geq \max_{k \in [2]} \mu_k$ , then we have that  $g_\mu(\theta_{m_1}, \theta_{m_2}) := \mu \log \left( \sum_{k=1}^2 \exp \left( \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \right) \right)$  is  $\frac{1}{\mu C_\mu}$ -PL.

*Proof.* Let  $\Theta = [\theta_{m_1}; \theta_{m_2}]$ . For brevity, we will denote  $\nabla_{\Theta} g(\Theta)$  as  $\nabla_{\Theta} g$ . First, note that we can have

$$\nabla_{\Theta} g = \frac{1}{\sum_{\ell=1}^2 \exp \left( \frac{f_{m_\ell}(\theta_{m_\ell}) - f_{m_\ell}^*}{\mu} \right)} \left[ \exp \left( \frac{f_{m_1}(\theta_{m_1}) - f_{m_1}^*}{\mu} \right) \nabla_{\theta_{m_1}} f_{m_1}, \exp \left( \frac{f_{m_2}(\theta_{m_2}) - f_{m_2}^*}{\mu} \right) \nabla_{\theta_{m_2}} f_{m_2} \right], \quad (35)$$

which in turn suggests that

$$\|\nabla_{\Theta} g\|^2 = \sum_{k=1}^2 \left( \frac{\exp\left(\frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu}\right)}{\sum_{\ell=1}^2 \exp\left(\frac{f_{m_\ell}(\theta_{m_\ell}) - f_{m_\ell}^*}{\mu}\right)} \right)^2 \|\nabla_{\theta_{m_k}} f_{m_k}\|^2. \quad (36)$$

Now consider any modality  $m_k$  for  $k \in [2]$ . From Assumption 4, we have that

$$\|\nabla_{\theta_{m_k}} f_{m_k}\|^2 \geq \frac{1}{\mu_k} (f_{m_k}(\theta_{m_k}) - f_{m_k}^*). \quad (37)$$

By choice of  $\mu = \max_{k \in [2]} \mu_k$ , we further have

$$\|\nabla_{\theta_{m_k}} f_{m_k}\|^2 \geq \frac{1}{\mu} (f_{m_k}(\theta_{m_k}) - f_{m_k}^*). \quad (38)$$

Taking the exponent of both sides of the above inequality and summing over  $k \in [2]$ , we get

$$\sum_{k=1}^2 \exp \|\nabla_{\theta_{m_k}} f_{m_k}\|^2 \geq \sum_{k=1}^2 \exp \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu} \geq \frac{1}{2} \sum_{k=1}^2 \exp \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu}. \quad (39)$$

Taking log of both sides of the above equation and using the sub-additivity on the right-hand side of the above inequality, we obtain

$$\sum_{k=1}^2 \|\nabla_{\theta_{m_k}} f_{m_k}\|^2 \geq \log \frac{1}{2} \sum_{k=1}^2 \exp \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu}. \quad (40)$$

Then, using the definition of  $C_\mu$ , we can have

$$C_\mu \sum_{k=1}^2 \left( \frac{\exp\left(\frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu}\right)}{\sum_{\ell=1}^2 \exp\left(\frac{f_{m_\ell}(\theta_{m_\ell}) - f_{m_\ell}^*}{\mu}\right)} \right)^2 \|\nabla_{\theta_{m_k}} f_{m_k}\|^2 \geq \log \frac{1}{2} \sum_{k=1}^2 \exp \frac{f_{m_k}(\theta_{m_k}) - f_{m_k}^*}{\mu}. \quad (41)$$

Dividing both sides of the above inequality by  $C_\mu$  will then give the desired result, i.e.

$$\|\nabla_{\Theta} g\|^2 \geq \frac{1}{\mu C_\mu} (g_\mu(\Theta) - g_\mu^*), \quad (42)$$

where  $g_\mu^* = \min_{\Theta} g_\mu(\Theta) = \mu \log 2$ .  $\square$

With the above Proposition, we can derive some insights on the optimality of the upper level multi-modal objective and lower level worst-performing uni-modal objective and the corresponding choice of hyperparameters, following theory given in (Shen and Chen, 2023).

**Theorem 2** ( (Shen and Chen, 2023) Proposition 2). *Let Assumptions 3-6 hold. Furthermore, let the choice of  $\mu$  be as suggested in Proposition 2. Then, for any  $\delta > 0$ , if we chose  $\lambda = \Theta \left( L_{1,mm} \sqrt{\frac{3\mu C_\mu}{\delta}} \right)$ , any local solution of 31 is a local solution of 30 with  $\epsilon \leq \delta$ .*

## E Related Work

**Multi-modal learning.** MML aims to process multi-sensory data for real-world tasks, with applications in fields such as sentiment classification (Zadeh et al., 2018; Cao et al., 2014), audio-visual localization (Tian et al., 2018), and visual question answering (Antol et al., 2015; Ilievski and Feng, 2017; Wu et al., 2021). Although integrating multiple modalities is expected to enhance performance, recent studies (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022; Li et al., 2023) reveal that the joint training paradigm often underutilizes modality-specific information. To address this, methods such as uncertainty awareness (Geng et al., 2021), gradient blending (Wang et al., 2020), learning rate adjustments (Wu et al., 2022; Yao and Mihalcea, 2022), and early stopping (Yao and

Mihalcea, 2022) have been proposed. More recent approaches adjust gradients based on output magnitudes for two modalities (Peng et al., 2022) or balance modality responses across any number of modalities (Li et al., 2023). (Wei and Hu, 2024) also apply the MOO method MGDA to mitigate imbalance issues in MML. In (Du et al., 2023), the authors propose two methods to exploit pre-trained uni-modal models to generate a multi-modal model. If multi-modal interactions are required in addition to uni-modal features learned in pre-training, the authors propose uni-modal teacher (UMT) to distill knowledge from uni-modal models while training a multi-modal model. If the pre-trained uni-modal models have learned strong features and no further cross-model interaction is required, (Du et al., 2023) propose a uni-modal ensemble (UME), which directly combines outputs of the pre-trained uni-modal models. These methods are in contrast to MIMO which does not require pre-trained uni-modal models and balances the learning of uni-modal features while training a multi-modal model. In (Zhang et al., 2023a), the authors propose a novel fusion technique called quality-aware multi-modal fusion (QMF), a dynamic method of fusion that assigns weight for uni-modal predictions based on "uncertainty" of the corresponding modality. A novel modality balancing method called ReconBoost is introduced in (Hua et al., 2024), where the goal is to update different modality specific learners in an alternating manner, such that the diversity of the updates is higher. (Javaloy et al., 2022) proposes a modality balancing method for training multi-modal VAEs that scales the gradient backpropagated from different modality heads according to the input dimension of each modality and uses the existing MTL-based method to calculate a conflict-averse gradient using the aforementioned gradients from modality heads. This method differs from MIMO in that (Javaloy et al., 2022) does not use separate uni-modal objectives to balance the modalities but controls the modality-specific gradients within the gradient of the multi-modal objective. In (Zhang et al., 2024), the authors propose a method called multi-modal learning with alternating uni-modal adaptation (MLA) to optimize each individual uni-modal encoder separately in an alternating manner and update the shared multi-modal head using recursive least squares algorithm to mitigate conflicts among updates corresponding to different modalities. Unlike these methods, we propose a simple gradient-based approach to address modality imbalance without requiring complex subroutines.

**Multi-objective optimization.** MOO focuses on optimizing multiple objectives simultaneously, often by balancing gradients across objectives. Usually, each objective corresponds to learning some "task". Common MOO approaches include task loss re-weighting based on uncertainty (Kendall et al., 2017), gradient norms (Chen et al., 2018), or task difficulty (Guo et al., 2018). Recent works (Désidéri, 2012; Sener and Koltun, 2018; Yu et al., 2020; Liu et al., 2021; Gu et al., 2021; Liu and Vicente, 2021; Zhou et al., 2022; Fernando et al., 2023; Chen et al., 2023) propose gradient aggregation to resolve task conflicts but at high computational cost. Alternatively, linear scalarization (Miettinen, 1999) simplifies MOO to single-objective optimization but may equally prioritize all objectives, which is suboptimal when preferences are unclear. Tchebyshev scalarization (Bowman Jr, 1976; Cortes et al., 2020; Lin et al., 2024) addresses this prioritizing worse performing objective, while lexicographic MOO (Miettinen, 1999; Guo et al., 2018) optimizes objectives in a pre-specified order. In this work, we build on the penalty reformulation method from (Shen and Chen, 2023) to reduce a bi-objective lexicographic problem to a single-objective problem and solve it using gradient-based optimization.

## F Additional Experiments and Details

In this section, we provide implementation details for MIMO and other baselines in CREMA-D, AV-MNIST, UR-FUNNY, and CMU-MOSEI MML benchmark datasets. Furthermore, we provide an ablation of MIMO parameters. For implementing uni-modal learning, vanilla MML, and balanced MML methods, we use the implementations of (Li et al., 2023)<sup>1</sup> and (Peng et al., 2022)<sup>2</sup>. MOO baselines are implemented by us. To be comparable with MOO methods, for uni-modal accuracy results for vanilla-MML and balanced MML methods, we train a dedicated uni-modal head using the features extracted from the uni-modal encoders, in addition to the multi-modal heads. We provide an average of over three seeds for our experiments, with an error bar of one standard deviation. All experiments are run using 2 NVIDIA GeForce RTX 3090 GPUs and 4 NVIDIA RTX A6000 GPUs.

**CREMA-D (Cao et al., 2014).** This dataset is for multi-modal speech emotion recognition using facial and vocal expressions. The dataset includes six common emotions: anger, happiness, sadness, neutrality, disgust, and fear. It is randomly divided into a training set with 6,027 samples, a validation set with 669 samples, and a

---

<sup>1</sup>[https://github.com/lihong2303/AGM\\_ICCV2023.git](https://github.com/lihong2303/AGM_ICCV2023.git)

<sup>2</sup>[https://github.com/GeWu-Lab/OGM-GE\\_CVPR2022](https://github.com/GeWu-Lab/OGM-GE_CVPR2022)

Table 6: Comparison using AV-MNIST dataset.

Method	Acc (%)	Acc <sub>a</sub> (%)	Acc <sub>v</sub> (%)	$t(s)$
$\mathcal{C}^a$	-	<b>42.32</b> $\pm 0.17$	-	0.019 $\pm 0.003$
$\mathcal{C}^v$	-	-	<b>65.05</b> $\pm 0.08$	0.019 $\pm 0.003$
<b>Vanilla MML</b>	71.70 $\pm 0.11$	39.98 $\pm 0.46$	64.67 $\pm 0.24$	0.015 $\pm 0.002$
<b>MSES</b>	71.61 $\pm 0.02$	39.92 $\pm 0.65$	64.66 $\pm 0.26$	0.017 $\pm 0.002$
<b>MSLR</b>	71.96 $\pm 0.12$	40.5 $\pm 0.79$	64.50 $\pm 0.14$	0.018 $\pm 0.002$
<b>OGM-GE</b>	71.70 $\pm 0.11$	39.98 $\pm 0.46$	64.67 $\pm 0.24$	0.055 $\pm 0.026$
<b>AGM</b>	70.92 $\pm 0.81$	29.16 $\pm 0.92$	61.76 $\pm 1.98$	0.060 $\pm 0.021$
<b>EW</b>	72.22 $\pm 0.04$	41.63 $\pm 0.26$	41.77 $\pm 0.18$	0.019 $\pm 0.011$
<b>MGDA</b>	72.15 $\pm 0.50$	41.63 $\pm 0.12$	41.92 $\pm 0.22$	0.086 $\pm 0.003$
<b>MMPareto</b>	72.42 $\pm 0.21$	41.64 $\pm 0.35$	41.83 $\pm 0.25$	0.085 $\pm 0.002$
<b>MIMO (ours)</b>	<b>72.77</b> $\pm 0.10$	42.21 $\pm 0.38$	42.25 $\pm 0.34$	<u>0.018</u> $\pm 0.004$

testing set with 745 samples. For method-specific parameter configurations for implementing uni-modal learning, vanilla MML, and balanced MML baselines we use the default setting of implementation by (Li et al., 2023). All methods are optimized with SGD optimizer with an initial stepsize of  $10^{-3}$ , for 100 epochs.

**UR-Funny (Hasan et al., 2019).** The dataset was created for affective computing tasks that detect humor through the use of words (text), gestures (vision), and prosodic cues (acoustic). This dataset was collected from TED talks and utilizes an equal number of binary labels for each sample. For method-specific parameter configurations for implementing uni-modal learning, vanilla MML, and balanced MML baselines we use the default setting of implementation by (Li et al., 2023). All methods are optimized with SGD optimizer with an initial stepsize of  $10^{-3}$ , for 100 epochs. Note that OGM-GE method is not implemented in this dataset since OGM-GE is by design only a two-modality balanced MML method.

**Kinetics-Sounds (Arandjelovic and Zisserman, 2017)** This dataset is derived from the larger Kinetics dataset Kay et al. (2017), which contains 400 classes of YouTube videos. Kinetics-Sounds specifically includes 31 human action categories that were selected for their potential to be both seen and heard, such as playing musical instruments. Each video clip is 10 seconds long, manually labeled for human actions via Mechanical Turk, and trimmed to center around the action of interest. The dataset comprises 19,000 video clips in total, with a split of 15,000 for training, 1,900 for validation, and 1,900 for testing.

**VGGSound (Chen et al., 2020)** This dataset is a comprehensive video dataset consisting of 309 classes that span a broad spectrum of audio events encountered in everyday contexts. The videos, each lasting 10 seconds, are recorded in real-world settings with an audio-visual alignment, meaning the source of the sound is visible. The dataset is partitioned following the original split in Chen et al. (2020). For our experiments, 168,618 videos are used for training and validation, while 13,954 are allocated for testing due to the unavailability of some YouTube videos.

**AV-MNIST (Vielzeuf et al., 2018).** In addition to the experiments given in the main text, here we provide a comparison of MIMO with proposed baselines in the AV-MNIST dataset. This dataset is for multi-media classification tasks by combining visual and audio features. The first modality, a noisy image, consists of  $28 \times 28$  PCA-projected MNIST images. The second modality, audio, consists of audio samples represented by  $112 \times 122$  spectrograms. The entire dataset comprises 70,000 samples, divided into a training set and a validation set at a ratio of 6 : 1. Additionally, 10% of the samples from both the training set and validation set were randomly selected to create a development set. For method-specific parameter configurations for implementing uni-modal learning, vanilla MML, and balanced MML baselines we use the default setting of implementation by (Li et al., 2023). All methods are optimized with SGD optimizer with an initial stepsize of  $10^{-3}$ , for 100 epochs. From Table 6, it can be seen that MIMO can outperform the best performing modality significantly, and perform comparably or better compared to other baselines. Moreover, when considering the subroutine execution times, MIMO is  $\sim 4$  faster compared to the next best performing method (MMPareto). These results demonstrate that MIMO can achieve superior performance with balanced MML, incurring only a minimal increase in computational time.

**CMU-MOSEI (Zadeh et al., 2018).** This dataset was compiled for sentence-level sentiment analysis and emotion recognition, consisting of 23,454 movie review clips drawn from over 65.9 hours of YouTube video featuring 1,000 speakers. As per the implementation in (Li et al., 2023), we utilize only the text and audio modalities, and the train/validation/test sets are split into 16,327, 1,871, and 4,662 samples, respectively. All methods are optimized with SGD optimizer with an initial stepsize of  $10^{-4}$ , for 100 epochs. The experiment results for CMU-MOSEI dataset is given in Table 7. It can be seen that while MIMO can outperform the best performing modality, vanilla MML and AGM fail to achieve this. Moreover, when considering the subroutine execution times, MIMO is only slightly slower than vanilla MML. These results demonstrate that MIMO can achieve superior performance with balanced MML, incurring only a minimal increase in computational time.

**AVE (Tian et al., 2018).** This dataset is an audio-visual dataset designed for event localization, encompassing 28 event classes. It contains 4,143 videos, each 10 seconds long, with synchronized audio and visual tracks, along with frame-level annotations. The experiment results for AVE dataset is given in Table 8. It can be seen that, similar to CMU-MOSEI dataset, MIMO can achieve superior performance compared to the baselines with a subroutine time similar to that of vanilla MML. Furthermore, it can be seen that MIMO can outperform unimodal baselines consistently, while vanilla MML and AGM fail to achieve this.

Table 7: Comparison using CMU-MOSEI dataset.

Method	Acc (%)	Acc <sub>t</sub> (%)	Acc <sub>a</sub> (%)	<i>t</i> (s)
Text	-	<b>81.53</b> ± 0.16	-	0.100± 0.009
Audio	-	-	74.12± 0.06	0.101± 0.009
<b>MML</b>	80.33± 0.18	73.89±1.58	73.08± 0.01	0.279± 0.009
<b>AGM</b>	80.28± 0.19	79.64±0.19	78.23± 0.65	0.304± 0.007
<b>MIMO</b>	<b>81.62</b> ± 0.06	81.44± 0.10	<b>81.36</b> ±0.23	0.287± 0.009

Table 8: Comparison using AVE dataset.

Method	Acc (%)	Acc <sub>a</sub> (%)	Acc <sub>v</sub> (%)	<i>t</i> (s)
Audio	-	66.03±0.28	-	0.010±0.001
Visual	-	-	63.82±0.99	0.011±0.001
<b>MML</b>	67.41±0.30	33.46±1.49	56.61±1.91	<b>0.027</b> ±0.002
<b>AGM</b>	72.54±1.13	54.73±1.29	50.92±1.98	0.161±0.002
<b>MIMO</b>	<b>73.69</b> ±0.24	<b>72.69</b> ±0.22	<b>71.85</b> ±0.47	0.029±0.002

**MOO baseline implementation.** We implement the equal weighting (EW) method by optimizing the sum of uni-modal and multi-modal objectives. For implementing MGDA, we consider the shared and non-shared parameters separately. Specifically, we solve the MGDA sub-problem (Fliege et al., 2019) using the gradient of uni-modal and multi-modal objectives with respect to encoder weights for each modality encoder. Non-shared parameters like multi-modal and uni-modal heads are updated using normal SGD updates. For MMPareto, we follow the method described in (Wei and Hu, 2024), with updating shared and non-shared parameters similar to that of MGDA implementation.

**Subroutine time calculation.** For calculating the subroutine times of MIMO and baselines, we compute the average computation time taken for the subroutine used for balancing modalities (if any) and updating the model parameters per batch. Since run times differ for different seeds due to background processes, we report the average subroutine times (over  $100 \times$  number of batches per epoch) calculated using one seed.

**MIMO parameters and implementation.** MIMO-specific parameters used for each dataset are given in Table 9. We coarsely tune  $\lambda$  parameter in the grid  $\{1, 10, 100\}$ , and  $\mu$  parameter in the grid  $\{0.001, 0.01, 0.1, 1.0\}$  for each dataset. To ensure numerical stability during MIMO implementation, when the loss values become large, we increase the value of  $\mu$  two times until the exponents in the MIMO objective fall within the permissible range for the datatype. The reported subroutine times include the computation time required for this adjustment.

Table 9: MIMO parameters

	CREMA-D	UR-Funny	Kinetics-Sound	VGGSound	AV-MNIST	CMU-MOSEI	AVE
$\lambda$	100	10	10	10	10	100	100
$\mu$	0.01	1.0	0.01	0.01	0.1	0.001	0.01