

Balancing Multi-modal Sensor Learning via Multi-objective Optimization

Heshan Fernando, Quan Xiao, Parikshit Ram, Yi Zhou, Horst Samulowitz, Nathalie Baracaldo, and Tianyi Chen

Abstract—Learning-enabled control systems increasingly rely on multiple sensing modalities (e.g., vision, audio, language, etc.) for perception and decision support. A key challenge is that multi-modal sensor training dynamics are often imbalanced: fast-to-learn sensing channels dominate optimization, while slower channels remain underutilized, degrading reliability under sensing perturbations. Existing balancing strategies are largely heuristic and can require computationally intensive subroutines. In this paper, we reformulate multi-modal sensor learning as a multi-objective optimization (MOO) problem that explicitly prioritizes the worst-performing modality while retaining the nominal multi-modal sensor fusion objective. We then propose a simple gradient-based method, MIMO (multi-modal sensor learning via MOO), for the resulting formulation. We provide convergence guarantees and evaluate the method on standard multi-modal benchmarks. Results show improved balanced performance over state-of-the-art balanced multi-modal learning and MOO baselines, together with up to $\sim 20\times$ reduction in subroutine computation time, highlighting the suitability of MIMO for resource-constrained control pipelines.

I. INTRODUCTION

Modern control and robotics systems routinely fuse heterogeneous sensing channels, including cameras, microphones, lidar/radar, and natural-language or operator signals. In these systems, multi-modal sensor learning (MSL) serves as an upstream estimation and representation module whose training dynamics strongly affect downstream control quality and robustness. Unlike uni-modal sensor pipelines, multi-modal sensor fusion learning can leverage complementary information across sensing channels, improving state inference and task performance under partial observability. This advantage has made multi-modal sensor fusion architectures central to embodied AI and autonomous decision-making stacks [1], [2], [3], [4].

A common sensor-fusion approach uses separate encoders for each sensor modality to transform data into feature representations that are then “fused” before further processing. The fused representation is subsequently processed by a model head to produce the output. In the two-modality case, the multi-modal sensor learning 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, θ_{m_1} and θ_{m_2} are encoders for modalities m_1 and m_2 , and ϑ_{mm} is the fusion head that

This work was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration. Heshan Fernando and Tianyi Chen are with Rensselaer Polytechnic Institute, Troy, NY, USA. Parikshit Ram, Yi Zhou, Horst Samulowitz, and Nathalie Baracaldo are with IBM Research, USA. Quan Xiao and Tianyi Chen are with Cornell University, Ithaca, NY, USA.

processes the fused features. The goal is then to optimize f_{mm} in (1) using standard optimization algorithms (e.g., SGD).

A. Imbalance issue in multi-modal sensor learning

While multi-modal sensor learning can outperform uni-modal sensor learning, recent studies show that standard training may still underperform the best single modality [5], [6]. For control-oriented systems, this issue is particularly critical: if one sensing stream dominates the updates, the learned estimator can become brittle to disturbances, occlusions, or sensor faults at deployment time. Prior works [7], [8], [9] attribute this behavior to dominant modalities suppressing slower but informative modalities during optimization. Using a toy example, Figure 1(a) illustrates this imbalance: θ_{m_1} and θ_{m_2} evolve at different rates, producing slower and biased convergence of the fused model.

Recent studies have provided theoretical explanations for this inefficiency by analyzing the training process of late-fusion models. In [10], [6], [11], the emergence of a dominant modality is explained through the concept of modality competition. In [12], the modality dominance is attributed to statistical characteristics of multi-modal data. Beyond these theoretical studies, empirical works have developed methods for balanced learning of different modalities, thereby improving performance [9], [13], [14], [9], [15], [16]. However, a principled algorithm with low per-iteration computational complexity for balanced multi-modal sensor learning, together with convergence guarantees, is still missing.

B. Our contributions

We address modality imbalance through a control-motivated optimization reformulation. The main idea is to shape training dynamics so that slow-to-learn sensing channels are not suppressed by dominant channels. We augment the fused objective with modality-specific objectives and enforce a lexicographic priority on the worst-performing uni-modal sensor channel. Concretely, we reformulate the multi-modal sensor learning problem in (1) as a lexicographic MOO problem [17], given by

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

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

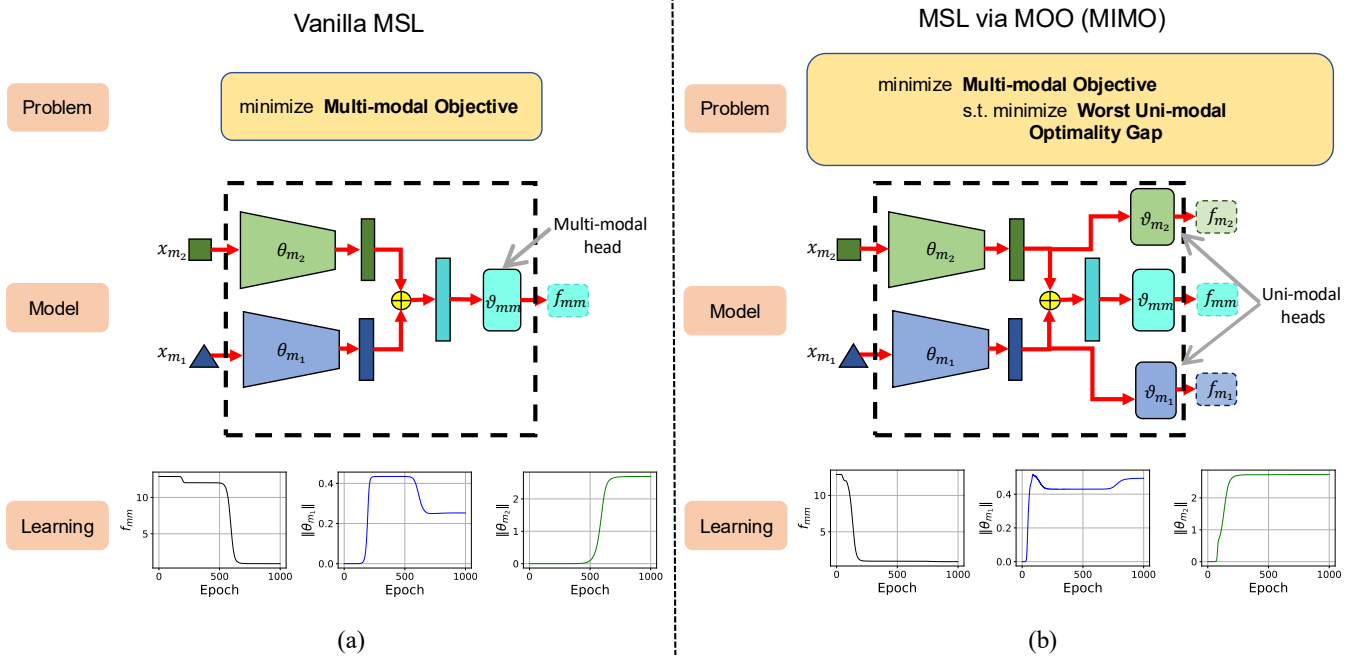


Fig. 1: **Balancing multi-modal sensor learning via multi-objective optimization.** (a) Optimizing the standard fusion objective can lead to slower convergence because fast-to-learn modalities dominate the optimization. (b) We propose MIMO (MSL via MOO), which optimizes a modified objective that prevents single-modality dominance, yielding faster convergence.

for all $k \in \{1, 2\}$. The formulation in (2) extends naturally to an arbitrary number of modalities in (8); here we focus on the two-modality case for clarity. We then propose MIMO (multi-modal sensor learning via MOO), a lightweight gradient-based algorithm for the resulting formulation. As shown in Figure 1(b), our method alleviates modality imbalance and converges quickly to the global optimum. Specifically, the weights of θ_{m_1} and θ_{m_2} are learned at similar rates (Figure 1(b), $\|\theta_{m_1}\|$ and $\|\theta_{m_2}\|$), leading to faster convergence.

Our main contributions are threefold. First, we introduce a lexicographic MOO reformulation of MSL that prioritizes worst-channel performance while preserving the original fused objective. Second, we develop MIMO, a simple first-order algorithm with low per-iteration overhead and no expensive inner-loop balancing subroutine. Third, we provide convergence guarantees and empirical evidence across standard benchmarks, showing improved balanced performance with up to $20\times$ lower subroutine time versus strong baselines.

II. PRELIMINARIES

In this section, we introduce basic concepts in multi-modal sensor learning [5] and MOO [17] used to build a formulation for balanced multi-modal sensor learning.

A. Multi-modal sensor learning (MSL)

Consider the classification problem on a multi-modal sensor dataset $\mathcal{D}_{mm} := \{x_i^{(m_1)}, x_i^{(m_2)}, \dots, x_i^{(m_K)}, y_i\}_{i=1}^N$, which consists of N input samples $x_i^{(m_k)}$ from K modalities m_k , where $k \in [K] := \{1, 2, \dots, K\}$, together with the corresponding labels y_i . Unlike the uni-modal sensor case, we need a “fusion” strategy to combine inputs from different

modalities before producing the output used in the loss function. Depending on whether fusion happens during feature extraction, after feature extraction, or in a hybrid manner, fusion strategies can be classified as *early*, *late*, or *hybrid fusion*, respectively [15]. In early fusion, data from different modalities are processed jointly starting from the raw inputs to obtain multi-modal features. In late fusion, separate encoders are used to extract uni-modal sensor features, which are then fused at the final stage of the model. Any combination of early and late fusion is referred to as hybrid fusion. We consider the late-fusion setting, where different modalities are fused after extracting features from each modality (see Figure 1(a)). Common fusion operations include summing the extracted features (summation) or concatenating them (concatenation) [9]. Let θ_{m_k} denote the parameters of the encoder that extracts features from inputs $x_i^{(m_k)}$ for all $k \in [K]$, and ϑ_{mm} denote the parameters of the fusion head that maps the fused features to the target output. Then, the problem of finding the optimal multi-modal sensor fusion model can be formulated 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 sensor fusion objective defined over all input modalities in $\mathcal{D}_{mm}$. Solving (3) yields the optimal fusion model Θ_{mm}^* .

Uni-modal bias in sensor-fusion training. As described in Section I, imbalance in sensor-fusion training can lead to poor fused performance, even relative to uni-modal sensor learning. To illustrate this issue, we use an example from [12] involving a multi-modal regression problem with a two-layer fully connected network, where one layer serves as the

encoder and the other serves as the fusion head, together with concatenation fusion. A toy implementation of this example is shown in Figure 1, and the corresponding implementation details are given in Appendix II.

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} (and similarly for modality m_2). Let the cross-correlation matrices between m_1 and m_2 be $C_{m_1m_2}$ and $C_{m_2m_1}$. The exact definitions of these matrices are given in Appendix II.

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

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

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

The decoupled nature of the output with respect to the modalities is a consequence of the late-fusion architecture with concatenation fusion. The corresponding regression loss is given by $f_{mm}(\Theta_{mm}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$. We can then derive the gradient of f_{mm} as follows (see Appendix II for details):

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

where

$$\begin{aligned} \Psi_1 &= C_{ym_1} - \vartheta_{mm,m_1} \theta_{m_1} C_{m_1} - \vartheta_{mm,m_2} \theta_{m_2} C_{m_2m_1}, \\ \Psi_2 &= C_{ym_2} - \vartheta_{mm,m_1} \theta_{m_1} C_{m_1m_2} - \vartheta_{mm,m_2} \theta_{m_2} C_{m_2}. \end{aligned}$$

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, for the model to have “learned” modality m_k , the weights corresponding to that modality should in general be nonzero, i.e., $\theta_{m_k} \neq 0$ and $\vartheta_{m_k} \neq 0$. Thus, to reach a stationary point while learning both modalities, we ideally 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. Since the model weights are initialized near zero, the model first visits a uni-modal sensor 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 achieves $\Psi_1 = \Psi_2 = 0$, $\vartheta_{mm,m_{3-k}} \neq 0$, and $\Theta_{m_{3-k}} \neq 0$ for all $k \in \{1, 2\}$. Which stationary point is visited first depends on the dataset statistics [12] (see Appendix II for more details). Furthermore, it can be shown that the uni-modal sensor stationary point visited first is decoupled from which modality contributes more to minimizing f_{mm} ; this phenomenon is known as “superficial modality preference” [12]. For example, in the toy MSL

task shown in Figure 1, the drop in the objective value f_{mm} is smaller when modality m_1 is learned (when the norm of the encoder weights $\|\theta_{m_1}\|$ attains a nonzero value) than when modality m_2 is learned, even though modality m_1 is learned first. Thus, imbalanced modality learning can lead to models that overfit a fast-learning modality and are therefore suboptimal for the overall multi-modal sensor fusion objective.

B. Lexicographic MOO

In this section, we introduce the key MOO formulation that we use to alleviate the imbalance issue in MSL discussed in Section II-A. Specifically, we introduce lexicographic MOO, which is useful when one can assign an *a priori* order to the learning objectives based on preference. In other words, optimizing lower-priority objectives is constrained by the optimality of higher-priority objectives. More concretely, consider a set of objectives $f_m : \mathbb{R}^d \mapsto \mathbb{R}$ for $m \in [M]$. Then, lexicographic MOO problem can be formulated as [17]

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

where the index of the objectives specifies the order in which the optimality of each objective should be achieved. For the bi-objective case, (6) can be written as the 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 us to incorporate prior knowledge about the problem into the optimization process. It can be shown that the solution of (6) is Pareto optimal [17].

III. BALANCED MSL VIA MOO

In this section, we first present our proposed reformulation of the multi-modal sensor learning problem to address the imbalance issue discussed in the previous section. We then detail the corresponding algorithm and provide a convergence analysis for the proposed method.

A. Problem formulation

In this section, we modify the original multi-modal sensor learning problem (3) to encourage balanced learning across modalities. Intuitively, the slow-to-learn modality will have the largest optimality gap. To alleviate this imbalance, we modify the MSL formulation to encourage learning of the modality with the worst optimality gap. Specifically, we seek to optimize the multi-modal sensor fusion objective subject to the optimality of the worst-performing (in terms of optimality gap) uni-modal sensor objective, namely,

$$\begin{aligned} &\min_{\Theta_{mm} \in \mathbb{R}^{d_{mm}}} f_{mm}(\Theta_{mm}) \\ &\text{s.t.} \quad \{\Theta_{m_k}\}_{k=1}^K \in \arg \min_{\{\Theta'_{m_k}\}_{k=1}^K} \max_{k \in [K]} (f_{m_k}(\Theta'_{m_k}) - f_{m_k}^*), \end{aligned} \quad (8)$$

where $\Theta_{m_k} = [\vartheta_{m_k}; \theta_{m_k}]$ with ϑ_{m_k} denoting the uni-modal sensor head dedicated to modality m_k , and $f_{m_k}^* =$

$\min_{\Theta_{m_k}} f_{m_k}(\Theta_{m_k})$ for all $k \in [K]$. Intuitively, formulation (8) learns the multi-modal sensor head by optimizing the fusion objective f_{mm} over the backbone model θ_{m_k} , which optimizes the worst-performing modality $\max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$.

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 sensor encoders) is shared between the two objectives, allowing for independent optimization of the non-shared part of $\hat{\Theta}_{mm}$ (multi- and uni-modal sensor heads).

The only parameters shared between the uni-modal sensor and multi-modal sensor fusion objectives are the uni-modal sensor encoders θ_{m_k} for all $k \in [K]$. Note that the lexicographic optimization of the worst-performing uni-modal sensor objective and the multi-modal sensor fusion objective with respect to the shared parameters θ_{m_k} for $k \in [K]$ can be viewed as a simple bilevel optimization problem. Recently, [18] introduced a reformulation of bilevel optimization as a single-level problem by penalizing lower-level constraints in the upper-level objective. Leveraging this view, we can obtain the single level MIMO objective by rewriting (8) as

$$\min_{\hat{\Theta}_{mm} \in \mathbb{R}^{d_{mm}}} \hat{f}_{mm}(\hat{\Theta}_{mm}) := f_{mm}(\Theta_{mm}) + \lambda \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*), \quad (9)$$

where $\hat{\Theta}_{mm} = [\vartheta_{mm}; \{\vartheta_{m_1}\}_{k=1}^K; \{\theta_{m_k}\}_{k=1}^K]$, and $\lambda > 0$ is a problem dependent penalty parameter.

How does 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, for the two modality case ($K = 2$), now augmented with additional uni-modal sensor heads (linear layers) to obtain uni-modal sensor objectives f_{m_1} and f_{m_2} (see Figure 1). Let the weight vectors corresponding to the uni-modal sensor heads be $\vartheta_{m_1} \in \mathbb{R}^{d_h}$ and $\vartheta_{m_2} \in \mathbb{R}^{d_h}$. Then, the new parameter vector can be written 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 sensor objectives are 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 $\hat{y}_i^{(m_k)} = \vartheta_{m_k}^\top \theta_{m_k}^\top x_i^{(m_k)}$ for each data index $i \in [N]$. With these definitions, we can rewrite the gradients of each layer of the network with additional uni-modal sensor heads as

$$\nabla_{\hat{\Theta}_{mm}} \hat{f}_{mm}(\hat{\Theta}_{mm}) = [\theta_{m_1}^\top (\Psi_1 + \lambda_1 \tilde{\Psi}_1); \vartheta_{mm, m_2}^\top (\Psi_2 + \lambda_2 \tilde{\Psi}_2); \Psi_1 \vartheta_{mm, m_1}^\top; \Psi_2 \vartheta_{mm, m_2}^\top; \lambda_1 \tilde{\Psi}_1 \theta_{m_1}^\top; \lambda_2 \tilde{\Psi}_2 \theta_{m_2}^\top], \quad (10)$$

where $\tilde{\Psi}_1 = C_{y m_1} - \vartheta_{m_1} \theta_{m_1} C_{m_1}$, $\tilde{\Psi}_2 = C_{y m_2} - \vartheta_{m_2} \theta_{m_2} C_{m_2}$, and $\lambda_i = \lambda$ if $i = \arg\max_{k \in \{1, 2\}} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$, and 0 otherwise. This modification, introduced by incorporating modality-specific objectives, helps balance multi-modal sensor learning in an intuitive way. Suppose modality m_1 is quick to learn. Then, initially, the weights in the m_2 component of the model remain close to zero, so optimality gap of m_2 is larger than that of m_1 . As a result, $\lambda_1 = 0$ and $\lambda_2 = \lambda$. Due to the

amplified gradient induced by the λ weighting, the weights in the m_2 component are updated more rapidly through the gradient contribution of the f_{m_2} objective, while modality m_1 is learned through the gradient contribution of f_{mm} . At stationarity, since all gradient components in (10) must be zero, the gradient components contributed by f_{mm} and f_{m_2} must each vanish. Hence, the model achieves optimality for the original multi-modal sensor fusion objective f_{mm} while simultaneously learning modality m_2 .

Formal justification of penalty reformulation. To provide more insights on why optimizing MIMO in (9) alleviates the imbalanced modality issue, we formally prove the approximate equivalence between the penalty reformulation (9) with the worst-case MOO problem (8).

Assumption 1 (PL condition of objectives). We assume that each modality objective $f_{m_k}(\Theta_{m_k})$ satisfies the Polyak-Łojasiewicz (PL) condition with μ , for all $k \in [K]$, i.e.

$$\|\nabla f_{m_k}(\Theta_{m_k})\|^2 \geq \mu(f_{m_k}(\Theta_{m_k}) - f_{m_k}^*).$$

PL condition is commonly used in bilevel optimization as a generalization of strongly convexity [18], [19], [20], [21], and holds for various non-convex applications such as linear quadratic regulator models [22], phase retrieval [23], over-parameterized neural network [24], including convolutional neural network [25], residual network [26], etc. Next, we introduce the notion of smooth functions.

Definition 1. 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

$$\|\nabla g(\Theta_1) - \nabla g(\Theta_2)\| \leq L\|\Theta_1 - \Theta_2\|. \quad (11)$$

Assumption 2 (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 $k \in [K]$.

Assumption 2 is standard in the optimization [27]. Under above assumptions, we can prove that the solutions of penalty problem (9) with certain penalty constant λ are also the solutions to the ϵ -relaxed worst-case problem (8) as follows:

$$\begin{aligned} \min_{\Theta_{mm} \in \mathbb{R}^{d_{mm}}} f_{mm}(\Theta_{mm}) \\ \text{s.t. } \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*) \leq \epsilon, \end{aligned} \quad (12)$$

Here, the constraint in (8) enforces that all modalities be learned equally well and requires the worst-case modality f_{m_k} to stay close to its best achievable performance $f_{m_k}^*$, thereby preventing it from being overlooked too much. We will then show that solving the MIMO problem (9) inherently addresses the relaxed worst-case problem (8).

Theorem 1 (Relations of the penalty reformulation and the relaxed problem). Under Assumption 1–2, let $(\Theta_{mm}^*, \Theta_{m_k}^*)$ be an ϵ_λ -global (or local) solution of the penalty problem (9) with λ . Then if $\lambda = \mathcal{O}(\epsilon_\lambda^{-1})$, then $(\Theta_{mm}^*, \Theta_{m_k}^*)$ is an ϵ_λ -global (or local) solution to the relaxed problem in (12), with some $\epsilon = \mathcal{O}(\epsilon_\lambda^2)$.

Algorithm 1 MIMO (multi-modal sensor learning via MOO)

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 λ , smoothing
parameter μ
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 $\hat{\Theta}_{mm,T+1}$

The proof of Theorem 1 is deferred to Appendix IV, which relies on the semi-smoothness and the nonsmooth PL property of $\max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$, and generalizes the results in [18] to the nonsmooth setting. Theorem 1 shows that solving the MIMO problem (9) is equivalent to solving the ϵ -relaxed worst-case problem (12), which motivates us to develop algorithm to solve MIMO problem (9) to alleviate the learning unbalanced issue for the worst-case modality.

B. Algorithm development

In this section, we present the algorithm used to solve the problem in (9). First, note that the penalty term in (9) is nonsmooth because of the max operator. In general, this kind of nonsmoothness leads to the slower convergence rate $\mathcal{O}(1/\epsilon^2)$ for subgradient-based optimization, compared to $\mathcal{O}(1/\epsilon)$ for smooth gradient-based optimization [28]. To alleviate this issue, prior work has proposed using smoothing functions [29] that encode structure in the nonsmooth term, yielding convergence to an $\mathcal{O}(\epsilon)$ -suboptimal point of the original nonsmooth problem within $\mathcal{O}(1/\epsilon)$ iterations. We next introduce the notion of smoothing functions.

Definition 2. 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_μ 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 μ , such that g_μ is $(L + \alpha\mu^{-1})$ -smooth (Definition 1).

Note that $g_\mu(\Theta) := \mu \log \left(\sum_{m=1}^M \exp(\mu^{-1} g_m(\Theta)) \right)$ is a smoothing function for $g(\Theta) = \max_{m \in [M]} g_m(\Theta)$ [29], where $\mu > 0$ is the smoothing parameter that controls the smoothness of g_μ . Then, the smoothed version of (9) is

$$\min_{\hat{\Theta}_{mm} \in \mathbb{R}^{\hat{d}_{mm}}} \hat{f}_{mm,\mu}(\hat{\Theta}_{mm}) := f_{mm}(\vartheta_{mm}, \theta_{m_1}, \theta_{m_2}) \quad (13)$$

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

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

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

where t is the iteration index and η_t is the learning rate. Specifically, for $K = 2$, the MIMO update in (14) can be

decomposed into updates for each component 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 $\sigma_{m_k} := \exp(h_{\mu,k}) / \sum_{k'=1}^2 \exp(h_{\mu,k'})$, $h_{\mu,k} = \mu^{-1} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*)$, and $k \in \{1, 2\}$. The MIMO algorithm is summarized in Algorithm 1. Importantly, the careful construction of a single objective for balanced multi-modal sensor learning allows us to use gradient descent directly on $\hat{f}_{mm,\mu}$, without computationally intensive heuristic subroutines that lack provable convergence guarantees. Next, we make the following assumptions to establish convergence of Algorithm 1.

Assumption 3 (Lipschitz continuity of objectives). For all $k \in [K]$, the 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 3 is required in our analysis to establish the smoothness of the composite objective $\hat{f}_{mm,\mu}$. We can then state the following proposition.

Proposition 1 (Smoothness of $\hat{f}_{mm,\mu}$). Under Assumptions 2 and 3, there exist $\hat{L}_{mm} > 0$ such that $\hat{f}_{mm,\mu}$ defined in (13) is $\hat{L}_{mm}$ -smooth (Definition 1), where $\hat{L}_{mm} := L_{mm} + \lambda \sum_{k=1}^2 (L_{m_k} + \mu^{-1} L_{m_k,1}^2)$.

A proof of Proposition 1 is given in Appendix III. Proposition 1 establishes the smoothness of the objective used by Algorithm 1 for gradient descent. Standard gradient descent theory [27] then yields the following result.

Theorem 2 (Convergence). Let Assumptions 2 and 3 hold. For any $\lambda, \mu > 0$, and $0 < \eta_t \leq 1/\hat{L}_{mm}$ for all $t \in [T]$, Algorithm 1 reaches an ϵ -stationary point of $\hat{f}_{mm,\mu}$ with iteration complexity $\mathcal{O}(1/\epsilon)$.

IV. EXPERIMENTS

In this section, we evaluate MIMO on several multi-modal benchmarks and compare it with popular multi-modal and MOO baselines. In addition to the performance comparison, we include experiments illustrating the generalization benefits of MIMO and an ablation study of its parameters.

Experiment settings. We adopt the same experimental settings as [15] and [9] across several popular multi-modal datasets. The **CREMA-D** dataset [30] is an audio-visual dataset for speech emotion recognition with six emotion labels. The **UR-Funny** dataset [31] was created for humor detection and involves word (text), gesture (vision), and intonational cue (acoustic) modalities. **Kinetics-Sound** [32] comprises 31 human action classes derived from the Kinetics dataset [33], which contains 400 categories of YouTube videos with both audio and visual components. **VGGSound** [34] is a large-scale video dataset with 309 classes spanning a broad range of everyday audio events. Additional experimental details and

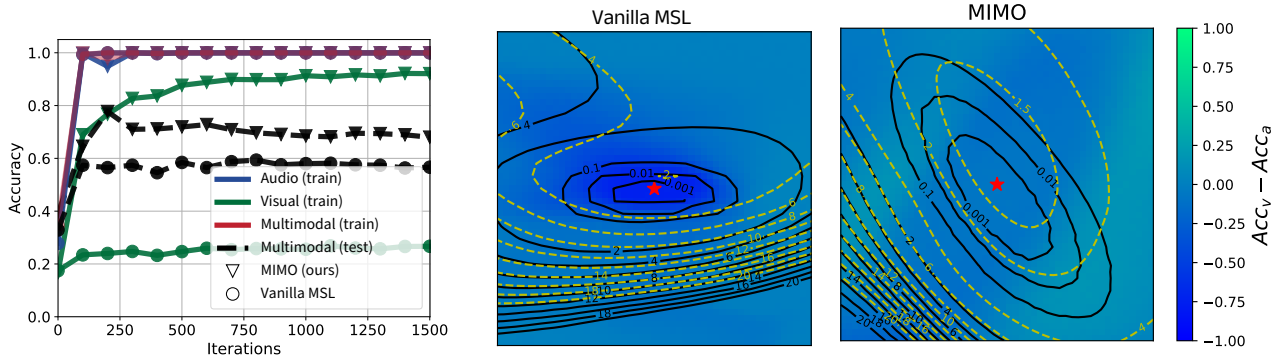


Fig. 2: **Left:** Comparison of the training and testing performance of MIMO algorithm with vanilla MSL (joint training with sum fusion) on CREMA-D dataset. **Middle and Right:** Comparison of the loss landscape of vanilla MSL and MIMO after 1500 iterations on CREMA-D dataset. The black contours (—) 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 sensor 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.

TABLE I: Multi-modal and uni-modal sensor test accuracy performance (Acc, Acc_a , Acc_v , Acc_t) of different MSL 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_a (%)	Acc_v (%)	$t(s)$	Acc (%)	Acc_a (%)	Acc_v (%)	Acc_t	$t(s)$
Audio	-	59.31 ± 0.76	-	0.028 ± 0.005	-	58.00 ± 0.74	-	-	0.037 ± 0.002
Visual	-	-	61.04 ± 0.87	0.029 ± 0.005	-	-	53.12 ± 0.52	-	0.038 ± 0.002
Text	-	-	-	-	-	-	-	63.74 ± 1.59	0.038 ± 0.002
MSL	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
MSES	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
MSLR	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
OGM-GE	74.49 ± 0.65	53.78 ± 1.21	47.82 ± 1.51	0.112 ± 0.009	-	-	-	-	-
AGM	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
EW	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
MGDA	63.47 ± 0.79	60.80 ± 0.68	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
MMPareto	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
MIMO	75.96 ± 0.83	55.60 ± 1.54	59.76 ± 1.40	<u>0.037</u> ± 0.012	64.54 ± 0.86	52.19 ± 1.08	50.38 ± 0.48	62.10 ± 1.31	<u>0.101</u> ± 0.013

results for AV-MNIST, AVE, and CMU-MOSEI are provided in Appendix V. We compare MIMO with several popular multi-modal baselines, including Modality-Specific Early Stopping (MSES) [13], Modality-Specific Learning Rate (MSLR) [14], On-the-fly Gradient Modulation Generalization Enhancement (OGM-GE) [9] (designed only for the two-modality case), and Adaptive Gradient Modulation (AGM) [15]. In addition to these baselines, we compare against MOO baselines such as equal weighting (EW), the Multiple Gradient Descent Algorithm (MGDA) [35], and MMPareto [16], which solve the multi-modal sensor learning problem as an MOO problem.

Balanced multi-modal sensor learning for better generalization. We first provide qualitative evidence that MIMO improves generalization in a real-world multi-modal sensor-fusion task using the CREMA-D dataset. Figure 2 (Left) shows the learning behavior of vanilla MSL and MIMO. While MIMO learns the slow-to-learn visual modality, vanilla

MSL overfits the audio modality, resulting in poor test performance. Furthermore, in Figure 2 (Middle and Right), we investigate the loss landscape around the models trained using vanilla MSL and the proposed MIMO method in a reduced-dimensional space [36]. Vanilla MSL achieves lower training loss (solid black contours) than MIMO, but exhibits poorer test loss (dashed yellow contours). In addition, the training-accuracy disparity between the audio and visual modalities (blue-green shading) is much larger in favor of the audio modality. In contrast, MIMO yields a more balanced training accuracy between the audio and visual modalities and a better test-loss profile than vanilla MSL, at the expense of slightly worse training loss. This suggests that balanced multi-modal sensor learning prevents the model from overfitting to a specific modality, thereby improving generalization.

Comparison with baselines. Next, we demonstrate the performance gains of MIMO on real-world multi-modal benchmarks relative to existing balanced multi-modal sensor

TABLE II: Comparison using VGGSound dataset.

	Acc (%)	Acc _a (%)	Acc _v (%)	$t(s)$
MSL	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	48.77 $\pm$ 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	69.10 $\pm$ 1.13	41.47 $\pm$ 1.04	38.20 $\pm$ 1.01	0.019 $\pm$ 0.004

TABLE III: Comparison using Kinetics-Sound dataset.

	Acc (%)	Acc _a (%)	Acc _v (%)	$t(s)$
MSL	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	0.026 $\pm$ 0.010
MMPareto	68.60 $\pm$ 1.41	48.07 $\pm$ 1.37	37.23 $\pm$ 1.37	0.689 $\pm$ 0.098
MIMO	69.60 $\pm$ 1.41	45.07 $\pm$ 1.04	43.13 $\pm$ 1.78	<u>0.039</u> $\pm$ 0.015

learning methods and MOO methods. Table I reports the performance of MIMO and several fusion and MOO baselines on the CREMA-D and UR-Funny classification benchmarks. On CREMA-D, MIMO achieves the best test accuracy among the baselines. Comparing against the individual-modality results also shows that MIMO attains superior performance through balanced multi-modal sensor learning, whereas vanilla MSL fails to match the best-performing individual modality (the visual modality). Furthermore, naively applying MOO methods to the fused and uni-modal sensor objectives (e.g., EW and MGDA) does not improve fused performance, because such approaches provide no fine-grained control over the trade-off between uni-modal sensor and fused-objective optimization. For example, MGDA is heavily biased toward the audio modality, leading to poor fused accuracy.

We then compare MIMO against MSL and MOO baselines on the three-modality UR-Funny benchmark and observe that MIMO performs comparably to or better than the baselines, while maintaining a subroutine time close to that of vanilla MSL; this is consistent with the observations on CREMA-D. Next, we evaluate performance on the Kinetics-Sound benchmark. The results in Table III show that MIMO outperforms both MOO baselines and balanced multi-modal sensor learning baselines. We attribute this superior performance to more balanced learning across modalities, as evidenced by the smaller disparity in uni-modal sensor accuracies for audio and visual modalities (only $\sim 3\%$ for MIMO, whereas the next-smallest disparity, achieved by MMPareto, is $\sim 11\%$). Similar observations can be made from the VGGSound results in Table II.

Furthermore, in most of the experiments above, MIMO has the fastest subroutine time (underlined values in the $t(s)$ column) among the top three best-performing methods, and its runtime is on the same order as vanilla MSL. In the VGGSound experiments, MIMO achieves a speed-up of up to $\sim 20\times$ relative to the next-best-performing baseline (MMPareto). While MIMO consistently outperforms the baselines across datasets, simple methods such as vanilla MSL can still work reasonably well on some datasets, such as UR-FUNNY. This may be due to the nature of the dataset, which does not satisfy the conditions for imbalance in MSL [37] and therefore

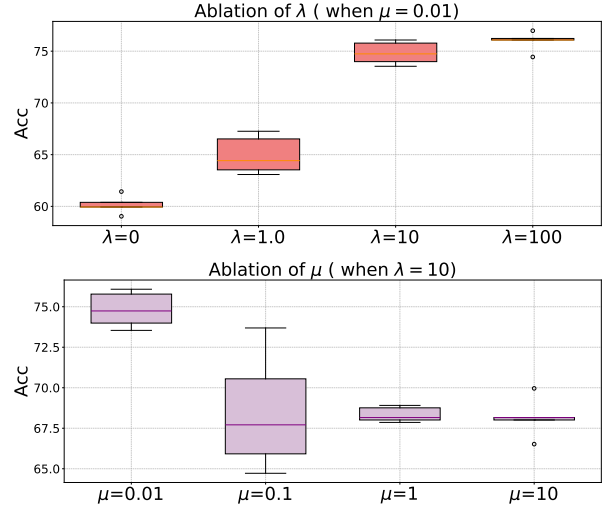


Fig. 3: Ablation of hyperparameters.

does not exhibit severe modality imbalance. However, since MIMO does not incur significant computational overhead, applying it in such cases is reasonable.

Ablation of MIMO parameters. In Figure 3, we provide an ablation study of the choice of λ and μ for MIMO on the CREMA-D dataset. We observe that, for very small λ ($= 1.0$), performance is worse than for larger λ values (≥ 10), which is expected because the effect of the constraint is weaker when λ is small. We also observe that increasing λ can improve performance, although the gains become marginal for larger λ values (≥ 10). For μ , smaller values lead to good performance, whereas larger values significantly degrade performance. This is also expected: when μ is large, the smoothed max deviates too much from the max function, preventing MIMO from prioritizing the worst-performing uni-modal sensor objective.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new formulation for balanced multi-modal sensor learning that prioritizes the worst-performing uni-modal sensor objective while optimizing the fused multi-modal sensor objective. This perspective improves modality balance and supports robust multi-modal sensor representations relevant to learning-enabled control pipelines. We then proposed MIMO, a simple gradient-based algorithm with convergence guarantees and without computationally expensive inner subroutines. Empirical evaluations show that MIMO outperforms existing balanced multi-modal sensor learning methods while substantially reducing computation time. This work focuses on late-fusion architectures with concatenation/sum fusion; extending MIMO-style MOO formulations to early/hybrid fusion and to explicit closed-loop control criteria is a natural next step.

REFERENCES

- [1] J. Wang, M. Gao, Y. Hu, R. R. Selvaraju, C. Ramaiah, R. Xu, J. F. JaJa, and L. S. Davis, "Tag: Boosting text-vqa via text-aware visual question-answer generation," *arXiv preprint arXiv:2208.01813*, 2022.

- [2] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *arXiv preprint arXiv:2010.03768*, 2020.
- [3] Z. Zhang, K. Chen, R. Wang, M. Utiyama, E. Sumita, Z. Li, and H. Zhao, "Neural machine translation with universal visual representation," in *International Conference on Learning Representations*, 2019.
- [4] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg *et al.*, "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.
- [5] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 695–12 705.
- [6] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably)," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9226–9259.
- [7] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.
- [8] P. Hu, X. Li, and Y. Zhou, "Shape: An unified approach to evaluate the contribution and cooperation of individual modalities," *arXiv preprint arXiv:2205.00302*, 2022.
- [9] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8238–8247.
- [10] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," *arXiv preprint arXiv:2012.09816*, 2020.
- [11] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 2551–2566, 2022.
- [12] Y. Zhang, P. E. Latham, and A. Saxe, "A theory of unimodal bias in multimodal learning," *arXiv preprint arXiv:2312.00935*, 2023.
- [13] N. Fujimori, R. Endo, Y. Kawai, and T. 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. Springer, 2020, pp. 412–422.
- [14] Y. Yao and R. Mihalcea, "Modality-specific learning rates for effective multimodal additive late-fusion," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1824–1834.
- [15] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 214–22 224.
- [16] Y. Wei and D. Hu, "Mmpareto: boosting multimodal learning with innocent unimodal assistance," in *International Conference on Machine Learning*, 2024.
- [17] K. Miettinen, *Nonlinear multiobjective optimization*. Springer Science & Business Media, 1999, vol. 12.
- [18] H. Shen and T. Chen, "On penalty-based bilevel gradient descent method," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 992–31 015.
- [19] Q. Xiao, S. Lu, and T. Chen, "A generalized alternating method for bilevel optimization under the polyak-łojasiewicz condition," in *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.
- [20] J. Kwon, D. Kwon, S. Wright, and R. Nowak, "On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation," in *Proc. International Conference on Learning Representations*, Vienna, Austria, 2024.
- [21] L. Chen, J. Xu, and J. Zhang, "On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis," in *Proc. of Conference on Learning Theory*. PMLR, 2024, pp. 947–980.
- [22] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. International Conference on Machine Learning*, Stockholm, Sweden, 2018.
- [23] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Foundations of Computational Mathematics*, vol. 18, pp. 1131–1198, 2018.
- [24] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Applied and Computational Harmonic Analysis*, vol. 59, pp. 85–116, 2022.
- [25] Y. Chen, Y. Shi, M. Dong, X. Yang, D. Li, Y. Wang, R. P. Dick, Q. Lv, Y. Zhao, F. Yang *et al.*, "Over-parameterized model optimization with Polyak-łojasiewicz condition," in *Proc. International Conference on Learning Representations*, Kigali, Rwanda, 2023.
- [26] P. Marion, Y.-H. Wu, M. E. Sander, and G. Biau, "Implicit regularization of deep residual networks towards neural odes," in *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.
- [27] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.
- [28] —, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, pp. 127–152, 2005.
- [29] X. Lin, X. Zhang, Z. Yang, F. Liu, Z. Wang, and Q. Zhang, "Smooth tchebycheff scalarization for multi-objective optimization," *arXiv preprint arXiv:2402.19078*, 2024.
- [30] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [31] M. K. Hasan, W. Rahman, A. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency *et al.*, "Ur-funny: A multimodal language dataset for understanding humor," *arXiv preprint arXiv:1904.06618*, 2019.
- [32] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 609–617.
- [33] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
- [34] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.
- [35] J.-A. Désidéri, "Multiple-gradient Descent Algorithm (MGDA) for Multi-objective Optimization," *Comptes Rendus Mathématique*, vol. 350, no. 5-6, 2012.
- [36] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," *Advances in neural information processing systems*, vol. 31, 2018.
- [37] Z. Lu, "A theory of multimodal learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [38] M. Ulbrich, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*. SIAM, 2011.
- [39] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.
- [40] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.
- [41] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. 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)*, 2018, pp. 2236–2246.
- [42] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 247–263.
- [43] J. Fliege, A. I. F. Vaz, and L. N. Vicente, "Complexity of Gradient Descent for Multi-objective Optimization," *Optimization Methods and Software*, vol. 34, no. 5, pp. 949–959, 2019.

Supplementary Material for “Balancing Multi-modal Sensor Learning via Multi-objective Optimization”

APPENDIX I NOTATIONS

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

TABLE IV: 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 sensor 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 MSL/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	
ϑ_{mm}	Multi-modal head in the vanilla MSL/MIMO model (see Figure 1)
ϑ_{m_k}	Uni-modal head for the modality m_k in the MIMO model (see Figure 1)
θ_{m_k}	Uni-modal encoder for the modality m_k in the vanilla MSL/MIMO model (see Figure 1)
Θ_{mm}	Concatenation of all the components in the vanilla MSL 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. $\hat{\Theta}_{mm} := [\vartheta_{mm}; \vartheta_{m_1}; \vartheta_{m_2}; \dots; \vartheta_{m_K}; \theta_{m_1}; \theta_{m_2}; \dots; \theta_{m_K}]$
Θ_{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 sensor fusion objective
$f_{m_k}(\Theta_{m_k})$	Uni-modal objective for modality m_k induced by uni-modal sensor head ϑ_{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	
ϑ_{mm, m_k}	Parameter matrix for the partition of ϑ_{mm} that corresponds to the modality m_k
C_{m_k}	Empirical input correlation matrix for modality m_k (defined in Section II)
$C_{m_k m_{3-k}}$	Empirical cross-correlation matrix between modality m_k and modality m_{3-k} (defined in Section II)
C_{ym_k}	Empirical input-output correlation matrix for modality m_k (defined in Section II)

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 $C_{m_2 m_1}$

APPENDIX II 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 [12].

a) Dataset.

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 μ and variance σ^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

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

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 [12] Appendix F.

TABLE V: Gradient of the objective for each layer of the network for vanilla MSL (column $f_{mm}(\Theta)$) and MIMO (column $f_{mm}(\Theta) + \text{column } \Delta$). Column Δ 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 ϑ_{m_1} and ϑ_{m_2} . In column Δ , $\lambda_i = \lambda$ if $i = \arg \max_{i \in \{1,2\}} (f_{m_k}(\vartheta_{m_k}, \theta_{m_k}) - f_{m_k}^*)$, else 0, where λ is the penalty parameter in (9).

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

b) Models.

For the vanilla multi-modal sensor fusion 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 fusion 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 sensor outputs (as shown in Figure 1 (b)).

c) Optimization.

We use a learning rate of 0.01 for both vanilla MSL and MIMO methods. For the MIMO method, we set $\lambda = 10$ and $\mu = 0.2$. The expressions of the gradients used in vanilla MSL and MIMO are summarized in Table V. The derivation of gradients follow a similar approach to that given in [12] Appendix A.

d) Superficial modality preference.

In Figure 1 (a), we can see that vanilla MSL is quick to learn modality m_1 . However, it does not contribute to minimizing the multi-modal sensor fusion objective compared to modality m_2 . This phenomenon is known as “superficial modality preference” [12], 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” [12] 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 Θ 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 sensor fusion objective f_{mm} parameterized by a late fusion multi-modal Θ 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.

APPENDIX III 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 Ψ 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 2), we have

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

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 3. 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 2), 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. $\square$

APPENDIX IV PROOF OF THEOREM 1

To prove Theorem 1, it is crucial to analyze the properties of the maximum gap function $g(\Theta) := \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$, where $\Theta_m = [\Theta_{m_1}, \dots, \Theta_{m_K}]$. As we will show later, although it is non-differentiable, it still has some desired properties.

Definition 3 (Semi-smoothness). *Let $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$ be a locally Lipschitz function. The function ℓ is said to be semi-smooth if it is directionally differentiable at any $z \in \mathbb{R}^d$ and for any direction $d \in \mathbb{R}^d$ and any $J \in \partial \ell(z + d)$, it holds that*

$$|\ell(z + d) - \ell(z) - J^\top d| = \mathcal{O}(\|d\|^2) \text{ as } d \rightarrow 0 \tag{30}$$

Furthermore, the function ℓ is said to be semi-smooth on $\mathcal{Z} \subset \mathbb{R}^d$ with L if ℓ is α -order semi-smooth at every $z \in \mathcal{Z}$ and there exists $\bar{d}$ such that for all $\|d\| \leq \bar{d}$ and any $z \in \mathcal{Z}$, the following holds

$$|\ell(z + d) - \ell(z) - J^\top d| \leq L \|d\|^2. \tag{31}$$

Both smooth and piecewise smooth functions are semi-smooth functions [38].

Proposition 2. *Under Assumption 1–2, the maximum gap function $g(\Theta_m) := \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*)$ satisfies the semi-smooth condition with $L := \max_{k \in [K]} L_{m_k}$ and the Kurdyka-Łojasiewicz (KL) condition, i.e.*

$$\text{dist}(0, \partial g(\Theta_m))^2 \geq \frac{\mu}{K} g(\Theta_m).$$

Proof. Since piecewise smooth function is semi-smooth [38], $\max_{k \in [K]} : \mathbb{R}^K \rightarrow \mathbb{R}$ is semi-smooth with modulus 1. Then since the class of semi-smooth functions is closed under composition [38], we know that $g(\Theta_m)$ is semi-smooth over Θ_m with $L := \max_{k \in [K]} L_{m_k}$.

Define the active index set as $\mathcal{I}(\Theta_m) = \arg \max_{k \in [K]} \{f_{m_k}(\Theta_{m_k}) - f_{m_k}^*\}$ and the block vector $\mathbf{G}^{(k)} = \mathbf{e}_k \otimes \nabla f_{m_k}(\Theta_{m_k})$, where $\mathbf{e}_k \in \mathbb{R}^K$ is the k -th standard basis vector. This means

$$\mathbf{G}^{(k)} = (0, \dots, 0, \nabla f_{m_k}(\Theta_{m_k}), 0, \dots, 0).$$

Then the subgradient of $g(\Theta_m)$ is given by [39]

$$\partial g(\Theta_m) = \text{conv}\{\mathbf{G}^{(k)} \mid k \in \mathcal{I}(\Theta_m)\}. \quad (32)$$

For any $v \in \partial g(\Theta_m)$, there exists $a_k \geq 0$ such that $\sum_{k \in \mathcal{I}(\Theta_m)} a_k = 1$ and k -th block of v has the form

$$v_k = \begin{cases} a_k \nabla f_{m_k}(\Theta_{m_k}), & k \in \mathcal{I}(\Theta_m), \\ 0 & k \notin \mathcal{I}(\Theta_m). \end{cases} \quad (33)$$

Therefore, we have

$$\begin{aligned} \|v\|^2 &= \sum_{k=1}^K \|v_k\|^2 = \sum_{k \in \mathcal{I}(\Theta_m)} \|a_k \nabla f_{m_k}(\Theta_{m_k})\|^2 \\ &\stackrel{(a)}{\geq} \mu \sum_{k \in \mathcal{I}(\Theta_m)} a_k^2 (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*) \\ &\stackrel{(b)}{=} \mu \sum_{k \in \mathcal{I}(\Theta_m)} a_k^2 g(\Theta_m) \\ &\stackrel{(c)}{\geq} \frac{1}{|\mathcal{I}(\Theta_m)|} \mu g(\Theta_m) \geq \frac{\mu}{K} g(\Theta_m) \end{aligned}$$

where (a) is because of Assumption 1, (b) is due to $k \in \mathcal{I}(\Theta_m)$, and (c) is because of the Cauchy–Schwarz inequality which gives $\sum_{k \in \mathcal{I}(\Theta_m)} a_k^2 \geq \frac{1}{|\mathcal{I}(\Theta_m)|}$. Taking inf over both sides for $v \in \partial g(\Theta_m)$, this completes the proof because $\text{dist}(0, \partial g(\Theta_m))^2 = \min_{v \in \partial g(\Theta_m)} \|v\|^2$. $\square$

Proof of Theorem 1. We only prove the local solution relations, and the corresponding global result then follows by extending the local region to the entire domain. Since $(\Theta_{mm}^*, \Theta_{m_k}^*)$ is the ϵ_λ -local solution of penalty problem (9), then there exists a local neighborhood of $(\Theta_{mm}^*, \Theta_{m_k}^*) \in \mathbb{B}((\Theta_{mm}^*, \Theta_{m_k}^*), R)$ with $R > 0$ such that for any $(\Theta_{mm}, \Theta_{m_k}) \in \mathbb{B}((\Theta_{mm}^*, \Theta_{m_k}^*), R)$, we have

$$f_{mm}(\Theta_{mm}) + \lambda \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*) \geq f_{mm}(\Theta_{mm}^*) + \lambda \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*) - \epsilon_\lambda \quad (34)$$

Specially, for any $(\Theta_{mm}, \Theta_{m_k}) \in \mathbb{B}((\Theta_{mm}^*, \Theta_{m_k}^*), R)$ with

$$\max_{k \in [K]} (f_{m_k}(\Theta_{m_k}) - f_{m_k}^*) \leq \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*) =: \epsilon$$

we have

$$f_{mm}(\Theta_{mm}) \geq f_{mm}(\Theta_{mm}^*) - \epsilon_\lambda \quad (35)$$

Therefore, $(\Theta_{mm}^*, \Theta_{m_k}^*)$ is ϵ_λ -local-optimal to the relaxed problem if $\epsilon := \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)$ is small.

Without loose of generality (W.L.O.G), we assume $0 \notin \nabla_{\Theta_{m_k}} f_{mm}(\Theta_{mm}^*) + \lambda \partial_{\Theta_{m_k}} \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)$. Denote $\hat{\Theta}_{mm}^* := [\Theta_{mm}^*, \{\Theta_{m_k}^*\}_{k=1}^K]$, and letting $\hat{\Theta}_{m_k}' = [\Theta_{mm}', \{\Theta_{m_k}'\}_{k=1}^K] = \hat{\Theta}_{mm}^* - \frac{ru}{\|u\|}$ where

$$u \in \nabla_{\Theta_{m_k}} f_{mm}(\Theta_{mm}^*) + \lambda \partial_{\Theta_{m_k}} \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)$$

is some subgradient that will be chosen later and $r \leq R$. Then according to Proposition 2 and Taylor expansion of f , we have

$$\begin{aligned} -\epsilon_\lambda &\leq f_{mm}(\Theta_{mm}') + \lambda \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}') - f_{m_k}^*) - (f_{mm}(\Theta_{mm}^*) + \lambda \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)) \\ &\leq u^\top (\Theta_{m_k}' - \Theta_{m_k}^*) + \frac{L_{mm} + \lambda \max_{k \in [K]} L_{m_k}}{2} \|\Theta_{m_k}' - \Theta_{m_k}^*\|^2 \leq -r\|u\| + \frac{L_{mm} + \lambda \max_{k \in [K]} L_{m_k}}{2} r^2 \end{aligned}$$

Therefore, when choosing $r = \mathcal{O}(R(\epsilon_\lambda/\lambda)^{\frac{1}{2}}) \leq R$, we have

$$\begin{aligned} & \text{dist}(0, \nabla_{\Theta_{m_k}} f_{mm}(\Theta_{mm}^*) + \lambda \partial_{\Theta_{m_k}} \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)) \\ & \leq \|u\| \leq \frac{L_{mm} + \lambda \max_{k \in [K]} L_{m_k}}{2} r + \frac{\epsilon_\lambda}{r} = \mathcal{O}(\lambda r + \epsilon_\lambda/r) = \mathcal{O}(\epsilon_\lambda^{\frac{1}{2}} \lambda^{\frac{1}{2}}) \end{aligned}$$

Let $u_g \in \partial_{\Theta_{m_k}} \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)$ be the one with $u = \nabla_{\Theta_{m_k}} f_{mm}(\Theta_{mm}^*) + \lambda u_g$, then by Cauchy-Schwartz inequality,

$$\text{dist}(0, \partial_{\Theta_{m_k}} \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*)) \leq \|u_g\| \leq \frac{\|u\|}{\lambda} + \frac{\|\nabla_{\Theta_{m_k}} f_{mm}(\Theta_{mm}^*)\|}{\lambda} \leq \mathcal{O}\left(\sqrt{\frac{\epsilon_\lambda}{\lambda}} + \frac{1}{\lambda}\right) = \mathcal{O}(\epsilon_\lambda).$$

where the last equality is obtained by setting $\lambda = \mathcal{O}(\epsilon_\lambda^{-1})$. Furthermore, with Proposition 2, we have

$$\frac{\mu}{K} \left(\max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*) \right) \leq \mathcal{O}(\epsilon_\lambda^2)$$

Therefore, we have $\epsilon := \max_{k \in [K]} (f_{m_k}(\Theta_{m_k}^*) - f_{m_k}^*) \leq \mathcal{O}(\epsilon_\lambda^2)$. Combining with (35), $(\Theta_{mm}^*, \Theta_{m_k}^*)$ is ϵ_λ -local-optimal to the relaxed problem with ϵ . $\square$

APPENDIX V 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 multi-modal sensor benchmark datasets. Furthermore, we provide an ablation of MIMO parameters. For implementing uni-modal sensor learning, vanilla MSL, and balanced multi-modal sensor learning methods, we use the implementations of [15]¹ and [9]². MOO baselines are implemented by us. To be comparable with MOO methods, for uni-modal sensor accuracy results for vanilla-MSL and balanced multi-modal sensor learning methods, we train a dedicated uni-modal sensor head using the features extracted from the uni-modal sensor encoders, in addition to the fusion 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.

a) CREMA-D [30].

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 testing set with 745 samples. For method-specific parameter configurations for implementing uni-modal sensor learning, vanilla MSL, and balanced multi-modal sensor learning baselines we use the default setting of implementation by [15]. All methods are optimized with SGD optimizer with an initial stepsize of 10^{-3} , for 100 epochs.

b) UR-Funny [31].

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 sensor learning, vanilla MSL, and balanced multi-modal sensor learning baselines we use the default setting of implementation by [15]. 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 multi-modal sensor learning method.

Kinetics-Sounds [32] This dataset is derived from the larger Kinetics dataset [33], 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 [34] 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 [34]. 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.

¹https://github.com/lihong2303/AGM_ICCV2023.git

²https://github.com/GeWu-Lab/OGM-GE_CVPR2022

TABLE VI: Comparison using AV-MNIST dataset.

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

c) AV-MNIST [40].

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 28×28 PCA-projected MNIST images. The second modality, audio, consists of audio samples represented by 112×122 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 sensor learning, vanilla MSL, and balanced multi-modal sensor learning baselines we use the default setting of implementation by [15]. All methods are optimized with SGD optimizer with an initial stepsize of 10^{-3} , for 100 epochs. From Table VI, 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 ~ 4 faster compared to the next best performing method (MMPareto). These results demonstrate that MIMO can achieve superior performance with balanced multi-modal sensor learning, incurring only a minimal increase in computational time.

d) CMU-MOSEI [41].

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 [15], 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 VII. It can be seen that while MIMO can outperform the best performing modality, vanilla MSL and AGM fail to achieve this. Moreover, when considering the subroutine execution times, MIMO is only slightly slower than vanilla MSL. These results demonstrate that MIMO can achieve superior performance with balanced multi-modal sensor learning, incurring only a minimal increase in computational time.

e) AVE [42].

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 VIII. 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 MSL. Furthermore, it can be seen that MIMO can outperform unimodal baselines consistently, while vanilla MSL and AGM fail to achieve this.

TABLE VII: Comparison using CMU-MOSEI dataset.

Method	Acc (%)	Acc _t (%)	Acc _a (%)	$t(s)$
Text	-	81.53 ± 0.16	-	0.100 ± 0.009
Audio	-	-	74.12 ± 0.06	0.101 ± 0.009
MSL	80.33 ± 0.18	73.89 ± 1.58	73.08 ± 0.01	0.279 ± 0.009
AGM	80.28 ± 0.19	79.64 ± 0.19	78.23 ± 0.65	0.304 ± 0.007
MIMO	81.62 ± 0.06	81.44 ± 0.10	81.36 ± 0.23	0.287 ± 0.009

f) MOO baseline implementation.

We implement the equal weighting (EW) method by optimizing the sum of uni-modal sensor and multi-modal sensor fusion objectives. For implementing MGDA, we consider the shared and non-shared parameters separately. Specifically, we solve the MGDA sub-problem [43] using the gradient of uni-modal sensor and multi-modal sensor fusion objectives with respect to encoder weights for each modality encoder. Non-shared parameters like multi-modal and uni-modal sensor heads are updated using normal SGD updates. For MMPareto, we follow the method described in [16], with updating shared and non-shared parameters similar to that of MGDA implementation.

TABLE VIII: Comparison using AVE dataset.

Method	Acc (%)	Acc _a (%)	Acc _v (%)	$t(s)$
Audio	-	66.03 ± 0.28	-	0.010 ± 0.001
Visual	-	-	63.82 ± 0.99	0.011 ± 0.001
MSL	67.41 ± 0.30	33.46 ± 1.49	56.61 ± 1.91	0.027 ± 0.002
AGM	72.54 ± 1.13	54.73 ± 1.29	50.92 ± 1.98	0.161 ± 0.002
MIMO	73.69 ± 0.24	72.69 ± 0.22	71.85 ± 0.47	0.029 ± 0.002

TABLE IX: MIMO parameters

	CREMA-D	UR-Funny	Kinetics-Sound	VGGSound	AV-MNIST	CMU-MOSEI	AVE
λ	100	10	10	10	10	100	100
μ	0.01	1.0	0.01	0.01	0.1	0.001	0.01

g) 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.

h) MIMO parameters and implementation.

MIMO-specific parameters used for each dataset are given in Table IX. We coarsely tune λ parameter in the grid $\{1, 10, 100\}$, and μ 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 μ 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.