MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance

Yake Wei [1] **Di Hu** [1 2]

Abstract

Multimodal learning methods with targeted unimodal learning objectives have exhibited their superior efficacy in alleviating the imbalanced multimodal learning problem. However, in this paper, we identify the previously ignored gradient conflict between multimodal and unimodal learning objectives, potentially misleading the unimodal encoder optimization. To well diminish these conflicts, we observe the discrepancy between multimodal loss and unimodal loss, where both gradient magnitude and covariance of the easier-to-learn multimodal loss are smaller than the unimodal one. With this property, we analyze Pareto integration under our multimodal scenario and propose MMPareto algorithm, which could ensure a final gradient with direction that is common to all learning objectives and enhanced magnitude to improve generalization, providing innocent unimodal assistance. Finally, experiments across multiple types of modalities and frameworks with dense cross-modal interaction indicate our superior and extendable method performance. Our method is also expected to facilitate multi-task cases with a clear discrepancy in task difficulty, demonstrating its ideal scalability. The source code and dataset are available at https://github.com/GeWu-Lab/MMPareto_ICML2024.

1. Introduction

People are immersed in a variety of sensors, encompassing sight, sound, and touch, which has sparked multimodal learning (Baltrušaitis et al., 2018; Lin et al., 2023), and audio-visual learning (Wei et al., 2022). Although many multimodal methods have revealed effectiveness, Peng et al. (2022) pointed out the imbalanced multimodal learning problem, where most multimodal models cannot jointly utilize all modalities well and the utilization of each modality is imbalanced. This problem has raised widely attention recently (Zhang et al., 2024). Several methods have been proposed to improve the training of worse learnt modality with additional module (Wang et al., 2020) or modality-specific training strategy (Peng et al., 2022; Wu et al., 2022; Wei et al., 2024). These methods often have one common sense that targetedly improves unimodal training. Among them, multitask-like methods that directly add unimodal learning objectives besides the multimodal joint learning objective, exhibit their superior effectiveness for alleviating this imbalanced multimodal learning problem (Wang et al., 2020; Du et al., 2023; Fan et al., 2023).

However, behind the effective performance, we observe a previously ignored risk in model optimization under this widely used multitask-like scenario, potentially limiting model ability. Every coin has two sides. Unimodal learning objectives undeniably effectively enhance the learning of corresponding modalities. Meanwhile, the optimization of parameters in unimodal encoder is influenced by both multimodal joint learning objective and its own unimodal learning objective. This entails the need to minimize two learning objectives concurrently, but usually, there does not exist a set of parameters that could satisfy this goal. Consequently, these multimodal and unimodal learning objectives could have conflict during optimization. In Figure 1a, we take an example of the video encoder on the widely used Kinetics Sounds dataset. As the results, negative cosine similarity indicates that multimodal and unimodal gradients indeed have conflicts in direction during optimization. Especially, these conflicts at the early training stage could substantially harm the model ability (Liu et al., 2020). Hence the primary multimodal learning is potentially disturbed.

To avoid optimization conflicts, it is essential to integrate both gradients well, making the unimodal gradient not affect the primary multimodal training but assist it. This necessity naturally accords with the idea of Pareto method (Sener & Koltun, 2018), which aims to find a steep gradient direction that benefits all objectives and finally converges to a trade-off state of them. As Figure 1b, on one of the representative multi-task cases, semantic segmentation and depth estimation tasks on Cityscapes (Cordts et al., 2016) dataset,

---

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Di Hu <dihu@ruc.edu.cn>.
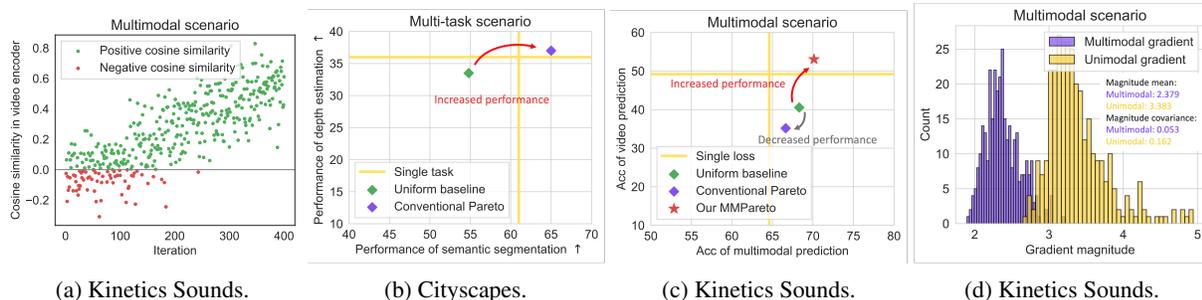
(a) Kinetics Sounds. (b) Cityscapes. (c) Kinetics Sounds. (d) Kinetics Sounds.

Figure 1. **(a).** Cosine similarity between multimodal and unimodal gradients in the video encoder of Kinetics Sounds dataset (Arandjelovic & Zisserman, 2017). **(b).** Methods performance on the multi-task dataset, Cityscapes (Cordts et al., 2016). Results are from (Sener & Koltun, 2018) **(c).** Methods performance of multimodal and unimodal prediction in the video encoder of Kinetics Sounds. Single loss is the result of the individually trained model with one corresponding learning objective. **(d).** The gradient magnitude distribution for a fixed video encoder of Kinetics Sounds dataset. Each count is a mini-batch of SGD optimization. Uniform baseline is a basic way where all losses are equally summed without special integration.

Pareto method has achieved ideal advancement in balancing the learning objective of these two tasks. Therefore, it is expected to keep superiority in solving conflicts in this multitask-like multimodal learning framework. However, the fact is contrary to the expectation. As Figure 1c, the conventional Pareto method (Sener & Koltun, 2018) surprisingly loses its efficacy, even is worse than the uniform baseline, where all gradients are equally summed.

To explore this counterintuitive phenomenon, we first analyze properties within multimodal learning scenarios. Unlike typical multi-task cases, multimodal joint loss is optimized with information from all modalities, while unimodal loss is optimized with information only from the corresponding modality. Accordingly, multimodal loss is naturally easier to learn, and verified with lower training error and faster convergence speed (Wang et al., 2020). We analyze the gradient magnitude distribution for a fixed set of parameters. As Figure 1d, the multimodal gradient magnitude is smaller than the unimodal one with a smaller batch sampling covariance. With these properties in gradient magnitude, we further theoretically analyze and empirically verify that the conventional Pareto could affect SGD noise strength and then bring the model to a sharper minima, weakening model generalization ability.

Based on the above analysis, it becomes imperative to well address gradient conflicts in multimodal scenarios. Hence, we propose the **M**ulti**M**odal **Pareto** (MMPareto) algorithm, which respectively takes the *direction* and *magnitude* into account during gradient integration. It ensures innocent unimodal assistance, where the final gradient is with direction common to all learning objectives and enhanced magnitude for improving generalization. We also provide an analysis of the method's convergence. Based on results across multiple types of datasets, our method effectively alleviates the imbalanced multimodal learning problem and could be well equipped with models with dense cross-modal interaction, like multimodal Transformers. As Figure 1c, our

method provides both advanced multimodal performance and unimodal performance. What's more, the unimodal performance is even superior to the individually trained unimodal model, which was rarely achieved before. Moreover, we verify that the proposed method could also extend to multi-task cases with clear discrepancy in task difficulty, indicating its scalability.

Our contribution is three-fold. **Firstly,** we observe the previously ignored gradient conflict in the widely used multitask-like framework for the imbalanced multimodal learning problem. **Secondly,** we theoretically analyze the failure of Pareto method in multimodal case, and then propose the MMPareto algorithm which could provide innocent unimodal assistance with enhanced generalization. **Thirdly**, experiments across different datasets verify our theoretical analysis as well as superior performance.

2. Related Work

2.1. Imbalanced Multimodal Learning

Recent research has uncovered the imbalanced multimodal learning problem, as multimodal models tend to favor specific modalities, thereby constraining their overall performance (Peng et al., 2022; Huang et al., 2022). Several methods have been proposed for this problem, with a shared focus on targetedly enhancing optimization of each modality (Wang et al., 2020; Peng et al., 2022; Wu et al., 2022; Fan et al., 2023). For example, Xu et al. (2023) focused on the fine-grained classification task, which has a higher demand for distinguishable feature distribution. Yang et al. (2024) put attention on the multimodal robustness. Wei et al. (2024) introduced a Shapley-based sample-level modality valuation metric, to observe and alleviate the fine-grained modality discrepancy. Among them, multitask-like methods that directly incorporate targeted unimodal constraints have demonstrated superior effectiveness (Wang et al., 2020; Du et al., 2023; Fan et al., 2023). However, under this multitask-

like framework, optimization of unimodal encoder is simultaneously controlled by the multimodal joint learning objective and corresponding unimodal learning objective, which could cause gradient conflict, potentially harming the primary multimodal learning. In this paper, we observe and diminish the potential conflict by the proposed MMPareto algorithm. Our method could effectively alleviate the imbalanced multimodal learning problem, achieving considerable improvement.

2.2. Pareto Integration in Multi-task Learning

Shared parameter in multi-task learning is expected to fit several learning objectives simultaneously, resulting in the potential conflict problem during optimization. Hence, the Pareto method is introduced to integrate different gradients, finding a gradient common to all objectives and finally converging to a trade-off state of them (Sener & Koltun, 2018). Besides, the idea of Pareto integration is extended from different perspectives, including more different trade-offs among different tasks (Lin et al., 2019) or faster convergence speed (Ma et al., 2020), to better benefit multi-task learning. Similarly, we observe the optimization conflict in the shared unimodal encoder in the multitask-like multimodal framework. Inspired by the success of Pareto integration, we introduce this idea but surprisingly find it failed. We further analyze and find the harmed generalization of Pareto integration in multimodal scenarios, and then propose MMPareto algorithm, which could handle multimodal scenarios and multi-task cases with clear discrepancy in task difficulty.

3. Method

3.1. Multitask-like Multimodal Framework

In multimodal learning, models are expected to produce correct predictions by integrating information from multiple modalities. Therefore, there are often *multimodal joint loss*, which takes the prediction of fused multimodal feature. However, only utilizing such joint loss to optimize all modalities together could result in the optimization process being dominated by one modality, leaving others being severely under-optimized (Peng et al., 2022; Huang et al., 2022). To overcome this imbalanced multimodal learning problem, introducing *unimodal loss* which targets the optimization of each modality is widely used and verified effective for alleviating this imbalanced multimodal learning problem (Wang et al., 2020). In these scenarios, the loss functions are:

$$\mathcal{L} = \mathcal{L}_m + \sum_{k=1}^{n} \mathcal{L}_u^k, \tag{1}$$

where $\mathcal{L}_m$ is the multimodal joint loss and $\mathcal{L}_u^k$ is the unimodal loss for modality k. n is the number of modalities. We mainly consider the multimodal discriminative task, and

all losses are cross-entropy loss functions. The illustration of this multitask-like multimodal framework is shown in the left part of Figure 2.

3.2. SGD Property and Hypothesis

In SGD optimization, based on former studies (Jastrzebski et al., 2017), for an arbitrary loss $\mathcal{L}$, when the batch size is sufficiently large, with the central limit theorem, the gradient of parameters θ at $t-$th mini-batch S, $\mathbf{g}_S(\theta(t)) = \frac{1}{|S|} \sum_{i=1}^{|S|} \nabla_{\theta(t)} \mathcal{L}(X_i, Y_i)$, is unbiased estimations of full gradient, $\mathbf{g}_N(\theta(t)) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta(t)} \mathcal{L}(X_i, Y_i)$:

$$\mathbf{g}_S(\theta(t)) \sim \mathcal{N}\left(\mathbf{g}_N(\theta(t)), \frac{1}{|S|}C\right). \tag{2}$$

In other words, $\mathbf{g}_S(\theta(t))$ is a random variable with covariance $\frac{1}{|S|}C$. C is brought by random batch sampling. N is the number of training samples. (X_i, Y_i) is a sample. The full gradient is calculated based on all training samples.

As Equation 1, multimodal framework has both multimodal loss function and unimodal loss function. For θ^k, the unimodal encoder parameter of modality k, gradients of $\mathcal{L}_m$ and $\mathcal{L}_u^k$ at iteration t satisfy:

$$\mathbf{g}_S^m(\theta^k(t)) \sim \mathcal{N}\left(\mathbf{g}_N^m(\theta^k(t)), \frac{1}{|S|}C^m\right), \tag{3}$$

$$\mathbf{g}_S^u(\theta^k(t)) \sim \mathcal{N}\left(\mathbf{g}_N^u(\theta^k(t)), \frac{1}{|S|}C^u\right), \tag{4}$$

where $\frac{1}{|S|}C^m$ and $\frac{1}{|S|}C^u$ are the batch sampling covariance.

In multimodal cases, unimodal loss only receives the prediction based on data of the corresponding modality. In contrast, the multimodal loss is optimized with more sufficient information from data of all modalities, making it easier to be trained. It has been verified that multimodal loss tends to converge with faster speed and with lower training error than unimodal one (Wang et al., 2020).

Beyond former studies, we also analyze the gradient magnitude distribution for a fixed set of parameters across different datasets. Results are as Figure 1d and Figure 4b. Each count is a different mini-batch. Firstly, multimodal gradients have a smaller magnitude than unimodal gradients, which could be brought by the lower multimodal error (Chen et al., 2018). Besides the magnitude, the batch sampling covariance $\frac{1}{|S|}C$ of gradient variables also differs between multimodal and unimodal ones. Multimodal gradients have a smaller covariance. The reason could be that more discriminative information from all modalities is more deterministic for classification, which aligns with the motivation of multimodal learning that introduces more modalities to reduce uncertainty (Trick et al., 2019). Then the multimodal training error has a less variants across different batches. Overall,
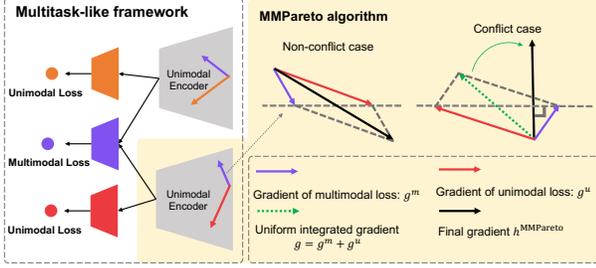
Figure 2. Illustration of multimodal framework and gradient integration strategy of our MMPareto.

based on former studies and our verification, we could conclude properties of multimodal and unimodal gradients, and have Hypothesis 1.

Hypothesis 1. *In multitask-like multimodal cases, for the shared unimodal encoder, the gradient of unimodal loss tends to have a larger magnitude and larger batch sampling covariance than easy-to-learnt multimodal loss.*

3.3. Pareto Integration in Multimodal Learning

3.3.1. PARETO INTEGRATION

In multimodal cases, multimodal loss and unimodal loss are tightly related, but their gradients may still exist conflicts, as Figure 1a. Hence, how to well integrate $\mathbf{g}_S^m(\theta^k(t))$ and $\mathbf{g}_S^u(\theta^k(t))$ needs to be solved. This accords with the idea of Pareto method in multi-task learning (Sener & Koltun, 2018). In Pareto method, at each iteration, gradients are assigned different weights, and the weighted combination is the final gradient, which can provide descent direction that benefits all learning objectives. Finally, parameters can converge to a trade-off state, Pareto-optimality, in which no objective can be advanced without harming any other objectives. It is natural to introduce Pareto integration into multimodal framework, avoiding conflict between multimodal and unimodal gradients. Concretely, in our case, for modality k, the Pareto algorithm is formulated to solve:

$$\min_{\alpha^m, \alpha^u \in \mathcal{R}} \|\alpha^m \mathbf{g}_S^m + \alpha^u \mathbf{g}_S^u\|^2 \tag{5}$$
$$s.t. \quad \alpha^m, \alpha^u \geq 0, \alpha^m + \alpha^u = 1,$$

where $\|\cdot\|$ denotes the L_2-norm. For brevity, we denote $\{\mathbf{g}_S^i(\theta^k(t))\}_{i\in\{m,u\}}$ as $\{\mathbf{g}_S^i\}_{i\in\{m,u\}}$ for modality k in some part. This problem is equal to finding the minimum-norm in the convex hull of the family of gradient vectors $\{\mathbf{g}_S^i\}_{i\in\{m,u\}}$. (Désidéri, 2012) showed that either found minimum-norm to this optimization problem is 0 and the corresponding parameters are Pareto-stationary which is a necessary condition for Pareto-optimality, or it can provide descent direction common to all learning objectives.

3.3.2. ANALYTIC SOLUTION OF PARETO

As results in Figure 1c, Pareto integration is expected to exhibit its advantage under multitask-like multimodal framework but surprisingly fails. To explore the hidden reason, we further analyze the property of the Pareto integration method. Based on Hypothesis 1, multimodal gradient tends to with a smaller magnitude than that of unimodal one, *i.e.,* $\|\mathbf{g}_S^m\| < \|\mathbf{g}_S^u\|$. Then, for the optimization problem of Equation 5, we can have its analytic solution:

$$\begin{cases} \alpha^m = 1, \alpha^u = 0 & \cos\beta \geq \frac{\|\mathbf{g}_S^m\|}{\|\mathbf{g}_S^u\|}, \\ \alpha^m = \frac{(\mathbf{g}_S^u - \mathbf{g}_S^m)^\top \mathbf{g}_S^u}{\|\mathbf{g}_S^m - \mathbf{g}_S^u\|^2}, \alpha^u = 1 - \alpha^m & \text{otherwise,} \end{cases}$$

where β is the angle between $\mathbf{g}_S^m$ and $\mathbf{g}_S^u$. Here we further analyze the above Pareto analytic solution. When $\cos\beta \geq \frac{\|\mathbf{g}_S^m\|}{\|\mathbf{g}_S^u\|}$, we have $\alpha^m > \alpha^u$. Otherwise, we also have:

$$\alpha^m - \alpha^u$$
$$= \frac{(\mathbf{g}_S^u - \mathbf{g}_S^m)^\top \mathbf{g}_S^u}{\|\mathbf{g}_S^m - \mathbf{g}_S^u\|^2} - (1 - \frac{(\mathbf{g}_S^u - \mathbf{g}_S^m)^\top \mathbf{g}_S^u}{\|\mathbf{g}_S^m - \mathbf{g}_S^u\|^2})$$
$$= \frac{\|\mathbf{g}_S^u\|^2 - \|\mathbf{g}_S^u\|\|\mathbf{g}_S^m\|\cos\beta}{\|\mathbf{g}_S^m - \mathbf{g}_S^u\|^2} - \frac{\|\mathbf{g}_S^m\|^2 - \|\mathbf{g}_S^u\|\|\mathbf{g}_S^m\|\cos\beta}{\|\mathbf{g}_S^m - \mathbf{g}_S^u\|^2}$$
$$> 0. \quad (\|\mathbf{g}_S^m\| < \|\mathbf{g}_S^u\|)$$

Remark 1. *The conventional Pareto method would assign a larger weight to the multimodal gradient with a smaller magnitude.*

Overall, as stated in Remark 1, we can conclude that the Pareto method tends to assign larger weight to the multimodal gradient $\mathbf{g}_S^m$ during integration, *i.e.,* $\alpha^m > \frac{1}{2}$.

3.3.3. GENERALIZATION HARMED RISK OF PARETO

In this section, we follow the notations of Section 3.2. During training, gradients of multiple losses are calculated separately, so they can be treated as being independent (Fan et al., 2022). Then, when without any gradient integration strategy *i.e.,* uniform baseline where all losses are equally summed, the final gradient is $\mathbf{h}_S(\theta^k(t)) = \mathbf{g}_S^m(\theta^k(t)) + \mathbf{g}_S^u(\theta^k(t))$. And based on Equation 3 and Equation 4, it satisfies:

$$\mathbf{h}_S(\theta^k(t)) \sim \mathcal{N}\left(\mathbf{g}_N^m(\theta^k(t)) + \mathbf{g}_N^u(\theta^k(t)), \frac{C^m + C^u}{|S|}\right). \tag{6}$$

Then, use the final gradient $\mathbf{h}_S(\theta^k(t))$ to update θ^k:

$$\theta^k(t+1) = \theta^k(t) - \eta\mathbf{h}_S(\theta^k(t)), \tag{7}$$
$$= \theta^k(t) - \eta\mathbf{h}_N(\theta^k(t)) + \eta\epsilon_t, \tag{8}$$

where $\mathbf{h}_N(\theta^k(t)) = \mathbf{g}_N^m(\theta^k(t)) + \mathbf{g}_N^u(\theta^k(t))$ and $\eta > 0$ is the learning rate. $\epsilon_t \sim \mathcal{N}\left(0, \frac{C^m + C^u}{|S|}\right)$ is often considered as the noise term of SGD (Zhu et al., 2018).

Then, we further consider the SGD optimization with conventional Pareto gradient integration. We know that at each iteration, Pareto method would return weight α^m and α^u for the integration of $\mathbf{g}_S^m(\theta^k(t))$ and $\mathbf{g}_S^u(\theta^k(t))$. Then, we have the final gradient after integration: $\mathbf{h}_S^{\text{Pareto}}(\theta^k(t)) = 2\alpha^m \mathbf{g}_S^m(\theta^k(t)) + 2\alpha^u \mathbf{g}_S^u(\theta^k(t))$, which satisfies:

$$\mathbf{h}_S^{\text{Pareto}}(\theta^k(t)) \sim \mathcal{N}\left(\mathbf{h}_N^{\text{Pareto}}(\theta^k(t)), \frac{(2\alpha^m)^2 C^m + (2\alpha^u)^2 C^u}{|S|}\right),$$
(9)

where $\mathbf{h}_N^{\text{Pareto}}(\theta^k(t)) = 2\alpha^m \mathbf{g}_N^m(\theta^k(t)) + 2\alpha^u \mathbf{g}_N^u(\theta^k(t))$. Here we use $2\alpha^i$ as the gradient weight to keep the same gradient weight summation with uniform baseline (*i.e.*, $2\alpha^m + 2\alpha^u = 2$). Then, when using Pareto integration, the parameter is updated as:

$$\theta^k(t+1) = \theta^k(t) - \eta \mathbf{h}_S^{\text{Pareto}}(\theta^k(t)) \qquad (10)$$

$$= \theta^k(t) - \eta \mathbf{h}_N^{\text{Pareto}}(\theta^k(t)) + \eta \zeta_t. \qquad (11)$$

$\zeta_t \sim \mathcal{N}\left(0, \frac{(2\alpha^m)^2 C^m + (2\alpha^u)^2 C^u}{|S|}\right)$ is the SGD noise term.

Based on Hypothesis 1, the unimodal gradient tends to have a larger batch sampling covariance. Suppose the covariance of multimodal gradient and unimodal gradient satisfies $kC^m = C^u$, where $k > 1$. Then, we can explore relation between ζ_t and ϵ_t. When the covariance of ζ_t is less than that of ϵ_t, it should satisfy:

$$(2\alpha^m)^2 C^m + (2\alpha^u)^2 C^u < C^m + C^u \qquad (12)$$

$$(2\alpha^m)^2 C^m + (2(1-\alpha^m))^2 \cdot kC^m < (k+1)C^m \qquad (13)$$

$$(2\alpha^m)^2 + 4k - 8\alpha^m k + (2\alpha^m)^2 k < k+1 \qquad (14)$$

$$\left(\alpha^m - \frac{1}{2}\right)\left((4k+4)\alpha^m + 2 - 6k\right) < 0 \qquad (15)$$

Hence when $\frac{1}{2} < \alpha^m < \frac{3k-1}{2k+2}$, the covariance of ζ_t is less than that of ϵ_t, *i.e.*, $\text{var}(\zeta_t) < \text{var}(\epsilon_t)$. In addition, based on Remark 1, Pareto weight satisfies $\frac{1}{2} < \alpha^m \leq 1$. Combine with the value range of α^m, then:

- For case $k \leq 3$, when $\frac{1}{2} < \alpha^m < \frac{3k-1}{2k+2}$, it has $\text{var}(\zeta_t) < \text{var}(\epsilon_t)$;

- For case $k > 3$, it has $\frac{3k-1}{2k+2} > 1$, $\text{var}(\zeta_t) < \text{var}(\epsilon_t)$ holds for all Pareto weight α^m.

Also, in Figure 3, we give the value of $\frac{3k-1}{2k+2}$ varies with k. When $k \leq 3$, it can observe a rapid increase of $\frac{3k-1}{2k+2}$ when the value of k increases. In addition, as shown in Figure 1d and Figure 4b, the difference of covariance between multimodal and unimodal gradient are often more than 3 times in experiments. These phenomena indicate that $\text{var}(\zeta_t) < \text{var}(\epsilon_t)$ could often happen in practice.

According to existing studies, the strength of noise term influences the minima found by SGD, and larger noise
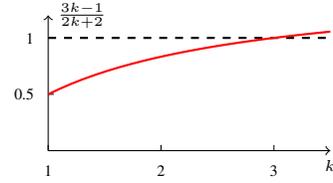

Figure 3. Value of $\frac{3k-1}{2k+2}$ varies with k.

term strength tends to bring flatter minima, which typically generalize well (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Neyshabur et al., 2017). Therefore, when $\text{var}(\zeta_t) < \text{var}(\epsilon_t)$, SGD noise term of conventional Pareto integration in Equation 11 can with smaller strength, compared with SGD noise term of uniform baseline in Equation 8. Then it would bring model with sharper minima, and further harm the model generalization ability.

3.4. Multimodal Pareto Algorithm

Based on the above analysis, conventional Pareto method could result in a sharper minima and then weakening model generalization in multimodal learning. It becomes essential to well address gradient conflicts in multimodal scenarios. In this paper, we propose the **MultiModal Pareto** (MMPareto) algorithm. It considers both the *direction* and *magnitude* during gradient integration, to provide innocent unimodal assistance where the final gradient is with a direction common to all learning objectives while an enhanced magnitude for improving generalization. Our method considers the conflict case and non-conflict case respectively. The overall algorithm is shown in Algorithm 1 and illustrated in Figure 2. In the following part, we take the encoder of modality k for an example, and encoders of all modalities follow the same integration. We also omit $\theta^k(t)$ for brevity.

Non-conflict case. We first consider the case $\cos\beta \geq 0$. Under this case, the cosine similarity between $\mathbf{g}_S^m$ and $\mathbf{g}_S^u$ is positive. For the direction, the arbitrary convex combination of the family of gradient vectors $\{\mathbf{g}_S^i\}_{i \in \{m,u\}}$ is common to all learning objectives. Therefore, we assign $2\alpha^m = 2\alpha^u = 1$ instead of analytic solution of Pareto during integration in this case, to have enhanced SGD noise term. With this setting, our final gradient $\mathbf{h}_S^{\text{MMPareto}} \sim \mathcal{N}\left(\mathbf{g}_N^m + \mathbf{g}_N^u, \frac{C^m + C^u}{|S|}\right)$. The noise term is $\xi_t \sim \mathcal{N}\left(0, \frac{C^m + C^u}{|S|}\right)$ with the enhanced strength, compared with conventional Pareto noise term ζ_t in Equation 11.

Conflict case. For the case $\cos\beta < 0$, it is essential to find the direction that is common to all losses and enhance the SGD noise strength during gradient integration. Hence we first solve the Pareto optimization problem of Equation 5, obtaining α^m and α^u, which could provide a non-conflict direction. Furthermore, to enhance the strength of noise term, we increase the magnitude of the final gradient. The

Algorithm 1 MMPareto

Require: Training dataset $\mathcal{D}$, iteration number T, initialized unimodal encoder parameters θ^k, $k \in \{1, 2, \cdots, n\}$, other parameters θ^{other}.

    for $t = 0, \cdots, T - 1$ **do**
        Sample a fresh mini-batch S from $\mathcal{D}$;
        Feed-forward the batched data S to the model;
        Calculate gradient using back-propagation;
        Update θ^{other} without gradient integration method;
        for $k = 1, \cdots, n$ **do**
            Obtain $\mathbf{g}_S^m$ and $\mathbf{g}_S^u$ for k-th unimodal encoder;
            Calculate $\cos \beta$; β is angle between $\mathbf{g}_S^m$ and $\mathbf{g}_S^u$;
            Solve problem of Equation 5, obtain α^m, α^u;
            if $\|\alpha^m \mathbf{g}_S^m + \alpha^u \mathbf{g}_S^u\| = 0$ **then**
                Find the Pareto stationarity;
            end if
            if $\cos \beta \geq 0$ **then**
                $2\alpha^m = 2\alpha^u = 1$;
            end if
            Integrate gradient: $\mathbf{h}_S' = 2\alpha^m \mathbf{g}_S^m + 2\alpha^u \mathbf{g}_S^u$;
            $\mathbf{h}_S^{\text{MMPareto}} = \underbrace{\mathbf{h}_S'/\|\mathbf{h}_S'\|}_{\text{Keep non-conflict direction}} \cdot \underbrace{\gamma\|\mathbf{g}_S^m + \mathbf{g}_S^u\|}_{\text{Enhanced magnitude}}$;
            Update θ^k with $\mathbf{h}_S^{\text{MMPareto}}$.
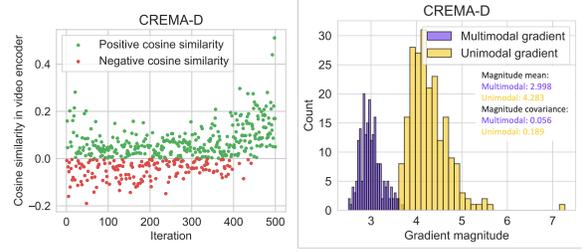   **end for**
    end for

magnitude of uniform baseline is used as the benchmark, to adjust in a proper range:

$$\mathbf{h}_S^{\text{MMPareto}} = \frac{2\alpha^m \mathbf{g}_S^m + 2\alpha^u \mathbf{g}_S^u}{\|2\alpha^m \mathbf{g}_S^m + 2\alpha^u \mathbf{g}_S^u\|} \cdot \|\mathbf{g}_S^m + \mathbf{g}_S^u\|. \quad (16)$$

In this case, the final gradient is $\mathbf{h}_S^{\text{MMPareto}} = \lambda \mathbf{h}_S^{\text{Pareto}}$, where $\lambda = \|\mathbf{g}_S^m + \mathbf{g}_S^u\|/\|2\alpha^m \mathbf{g}_S^m + 2\alpha^u \mathbf{g}_S^u\|$. Based on Remark 1, the smaller multimodal magnitude is with larger weight, we can have that $\|2\alpha^m \mathbf{g}_S^m + 2\alpha^u \mathbf{g}_S^u\| < \|\mathbf{g}_S^m + \mathbf{g}_S^u\|$, and accordingly $\lambda > 1$. Then, the final gradient satisfies: $\mathbf{h}_S^{\text{MMPareto}} \sim \mathcal{N}\left(\lambda \mathbf{h}_N^{\text{Pareto}}, \lambda^2 \cdot \frac{(2\alpha^m)^2 C^m + (2\alpha^u)^2 C^u}{|S|}\right)$. The noise term is $\xi_t \sim \mathcal{N}\left(0, \lambda^2 \cdot \frac{(2\alpha^m)^2 C^m + (2\alpha^u)^2 C^u}{|S|}\right)$, which has a larger strength than noise term ζ_t of conventional Pareto in Equation 11. The noise strength is enhanced, and model generalization is accordingly improved.

After integration, to further enhance the generalization, we slightly increase the magnitude of the final gradient: $\mathbf{h}_S^{\text{MMPareto}} = \gamma \mathbf{h}_S^{\text{MMPareto}}$, where $\gamma > 1$ is the rescale factor.

Overall, MMPareto provides both non-conflict direction and enhanced SGD noise strength, helping the model converge to a flatter minima and generalize better. Beyond that, we also analyze the convergence of proposed MMPareto method. As Remark 2, our algorithm is guaranteed to converge to a Pareto stationarity. Detailed proof and related experiments are provided in Appendix E and Appendix F.

(a) Direction conflict.   (b) Gradient distribution.

Figure 4. **(a).** Cosine similarity between gradients of multimodal and unimodal loss in the video encoder of CREMA-D. **(b).** Gradient magnitude distribution in the video encoder of CREMA-D.

Remark 2. *The proposed MMPareto method admits an iteration sequence that converges to a Pareto stationarity.*

4. Experiment

4.1. Dataset and Experiment Settings

CREMA-D (Cao et al., 2014) is an audio-visual dataset for emotion recognition, covering 6 usual emotions. **Kinetics Sounds** (Arandjelovic & Zisserman, 2017) is an audio-visual dataset containing 31 human action classes. **Colored-and-gray-MNIST** (Kim et al., 2019) is a synthetic dataset based on MNIST (LeCun et al., 1998). Each instance contains two kinds of images, a gray-scale and a monochromatic colored image. **ModelNet40** (Wu et al., 2015) is a dataset with 3D objects, covering 40 categories. This dataset could be used to classify 3D objects based on the multiple 2D views data (Su et al., 2015).

When not specified, ResNet-18 (He et al., 2016) is used as the backbone in experiments and models are trained from scratch. Unimodal modal features are integrated with late fusion method. Specifically, for the Colored-and-gray MNIST dataset, we build a neural network with 4 convolution layers and 1 average pool layer as the encoder, like (Fan et al., 2023) does. During the training, we use SGD with momentum (0.9) and $\gamma = 1.5$ in experiments. More details are provided in Appendix B.

4.2. Gradient Conflict and Magnitude Distribution in Multimodal Scenarios

Here we verify the direction conflict and magnitude distribution between multimodal and unimodal gradient across different datasets. Firstly, in Figure 1a and Figure 4a, we show the cosine similarity between gradients on the Kinetics Sounds and CREMA-D. Based on the results, the update direction of multimodal and unimodal gradient indeed have conflict, *i.e.,* negative cosine similarity, which potentially brings risk for the optimization of the corresponding shared unimodal encoder. In addition, such conflicts often exist in the early training stage, disturbance in which stage could substantially harm the model ability (Liu et al., 2020). In
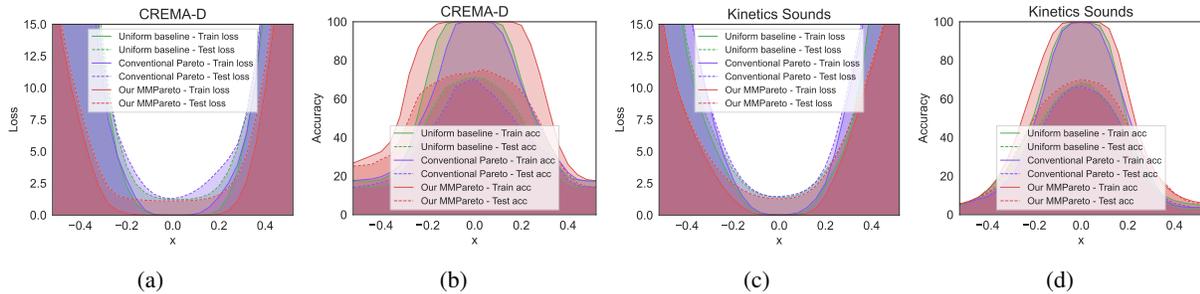
Figure 5. Visualization of the loss landscape and corresponding accuracy of uniform baseline, conventional Pareto and our MMPareto methods. Our MMPareto method brings flatter minima. The visualization method is from (Li et al., 2018). Uniform baseline is a basic way where all losses are equally summed without special integration.

Table 1. **Comparison with imbalanced multimodal learning methods where bold and underline represent the best and runner-up respectively.** * indicates that the unimodal evaluation (Acc audio and Acc video) is obtained by fine-tuning a unimodal classifier with frozen trained unimodal encoder, since this method could not provide unimodal prediction directly. This evaluation method borrows from (Peng et al., 2022). (↓) indicates a performance drop compared with uniform baseline where all losses are equally summed.

Method	CREMA-D			Kinetics Sounds		
	Acc	Acc audio	Acc video	Acc	Acc audio	Acc video
Audio-only	-	61.69	-	-	<u>53.63</u>	-
Video-only	-	-	<u>56.05</u>	-	-	49.20
Unimodal pre-trained & fine-tune	71.51	60.08	**60.22**	68.75	53.49	<u>50.07</u>
One joint loss*	66.13	59.27	36.56	64.61	52.03	35.47
Uniform baseline	71.10	<u>63.44</u>	51.34	68.31	53.20	40.55
G-Blending (Wang et al., 2020)	<u>72.01</u>	60.62 (↓)	52.23	<u>68.90</u>	52.11 (↓)	41.35
OGM (Peng et al., 2022)*	69.19 (↓)	56.99 (↓)	40.05 (↓)	66.79 (↓)	51.09 (↓)	37.86 (↓)
Greedy (Wu et al., 2022)*	67.61 (↓)	60.69 (↓)	38.17 (↓)	65.32 (↓)	50.58 (↓)	35.97 (↓)
PMR (Fan et al., 2023)*	66.32 (↓)	59.95 (↓)	32.53 (↓)	65.70 (↓)	52.47 (↓)	34.52 (↓)
AGM (Li et al., 2023)*	70.06 (↓)	60.38 (↓)	37.54 (↓)	66.17 (↓)	51.31 (↓)	34.83 (↓)
MMPareto	**75.13**	**65.46**	55.24	**70.13**	**56.40**	**53.05**

addition, as Figure 1d and Figure 4b, we also observe the gradient magnitude distribution for a fixed set of parameters across different datasets. Each count is a different mini-batch. Based on the results, we could conclude that the multimodal gradient has a smaller magnitude and smaller batch sampling covariance, compared with the unimodal one. The difference in covariance could even be more than 3 times, greatly affecting model generalization.

4.3. Loss Landscape Analysis

Based on our analysis, conventional Pareto method could result in a sharper minima and then weakening model generalization. To empirically verify it, we visualize the loss landscape and the corresponding accuracy of uniform baseline, conventional Pareto method, and our MMPareto method. As shown in Figure 5, conventional Pareto method indeed has a sharper minima, compared with uniform baseline where multimodal and unimodal losses are equally summed. Its loss and accuracy change more drastically. What's more, loss landscape near the minima that our MMPareto method converges on is the flattest one, since it ensures gradient

with both non-conflict direction and enhanced SGD noise strength during optimization.

In addition, as shown in Figure 5, the visually flatter minima consistently correspond to a lower test error. For example, our MMPareto method with flatter minima has a lower test error and higher accuracy. This phenomenon verifies that the flatter minima tend to generalize better, supporting our former analysis.

4.4. Comparison with Related Imbalanced Methods

To validate the effectiveness of our MMPareto method in overcoming imbalanced multimodal learning problems, we compare it with recent studies: G-Blending (Wang et al., 2020), OGM-GE (Peng et al., 2022), Greedy (Wu et al., 2022), PMR (Fan et al., 2023) and AGM (Li et al., 2023).

In addition, we also compare several basic settings. **Audio- and Video-only** are unimodal models trained individually. **Unimodal pre-trained & fine-tune** is the method that fine-tunes the multimodal model with individually pre-trained unimodal encoders. **One joint loss** is the method that

Table 2. **Comparison with imbalanced multimodal learning methods where bold and underline represent the best and runner-up respectively.** The network is transformer-based framework, MBT (Nagrani et al., 2021).

Method	CREMA-D				Kinetics Sounds			
	from scratch		with pretrain		from scratch		with pretrain	
	Acc	macro F1	Acc	macro F1	Acc	macro F1	Acc	macro F1
One joint loss	44.96	42.78	66.69	67.26	42.51	41.56	68.30	69.31
Uniform baseline	45.30	43.74	69.89	<u>70.11</u>	43.31	43.08	69.40	69.60
G-Blending (Wang et al., 2020)	<u>46.38</u>	<u>45.16</u>	<u>69.91</u>	70.01	<u>44.69</u>	<u>44.19</u>	69.41	69.47
OGM-GE (Peng et al., 2022)	42.88	39.34	65.73	65.88	41.79	41.09	69.55	69.53
Greedy (Wu et al., 2022)	44.49	42.76	66.67	67.26	43.31	43.08	69.62	69.75
PMR (Fan et al., 2023)	44.76	42.95	65.59	66.07	43.75	43.21	<u>69.67</u>	<u>69.87</u>
AGM (Li et al., 2023)	45.36	43.81	66.54	67.75	43.65	43.57	69.59	69.14
MMPareto	**48.66**	**48.17**	**70.43**	**71.17**	**45.20**	**45.26**	**70.28**	**70.11**

Table 3. **Comparison with related multi-task methods on Colored-and-gray-MNIST, ModelNet40 and Kinetics Sounds.** Bold and underline represent the best and runner-up respectively.

Method	CG-MNIST		ModelNet40		Kinetics Sounds	
	Acc	macro F1	Acc	macro F1	Acc	macro F1
One joint loss	60.50	59.89	87.88	83.32	64.61	64.12
Uniform baseline	75.68	75.66	89.18	84.69	68.31	68.13
Conventional Pareto (Sener & Koltun, 2018)	62.00	61.85	88.05	83.01	66.64	66.17
GradNorm (Chen et al., 2018)	76.16	76.12	88.98	83.79	65.84	65.14
PCGrad (Yu et al., 2020)	<u>79.35</u>	77.14	89.59	84.44	<u>69.11</u>	<u>68.75</u>
MetaBalance (He et al., 2022)	79.18	<u>77.87</u>	<u>89.63</u>	<u>84.87</u>	68.90	68.62
MMPareto	**81.88**	**81.69**	**89.95**	**85.15**	**70.13**	**70.18**

only uses multimodal joint loss. And **uniform baseline** is the method in which multimodal and unimodal losses are equally summed. To comprehensively evaluate the model ability, we further observe the unimodal performance, besides the common multimodal performance.

Based on Table 1, we can find that the uniform baseline can achieve considerable performance, and even could outperform or be comparable with these imbalanced multimodal learning methods. The reason could be that the introduction of unimodal loss effectively enhances the learning of each modality, which accords with the core idea of these compared methods. Moreover, our MMpareto method with a conflict-free optimization process achieves a considerable improvement, compared with existing methods at the multimodal prediction. More than that, our MMPareto method simultaneously exhibits outstanding unimodal performance, and even can outperform individually trained unimodal model. For example, Audio accuracy of MMPareto is superior to Audio-only method on both CREMA-D and Kinetics Sounds datasets. This was rarely achieved in before studies.

Besides, we also conduct experiments under the widely used Transformer backbone, MBT (Nagrani et al., 2021), which contains both single-modal layers and cross-modal interaction layers. Compared to the former CNN backbone with the late fusion method, unimodal features in this transformer-based framework are more fully interacted and integrated. During experiments, we conduct experiments both from scratch and with ImageNet pre-training. Results are shown in Table 2. Based on the results, we can have the following observation. Firstly, former imbalanced multimodal learning could lose efficacy under these more complex scenarios with cross-modal interaction. For example, OGM-GE method is even worse than the one joint loss method on CREMA-D dataset. In contrast, our MMPareto gradient integration strategy is not only applicable to CNN backbones, but also able to maintain superior performance in transformer-based frameworks with complex interactions. In addition, whether or not to use pre-training does not affect the effectiveness of our method, which reflects its flexibility.

4.5. Comparison with Related Multi-task Methods

In past studies, there are other strategies that are used to balance multiple learning objectives. In this section, besides conventional Pareto (Sener & Koltun, 2018), we also compare several representative ones: GradNorm (Chen et al., 2018), PCGrad (Yu et al., 2020), MetaBalance (He et al., 2022). Experiments are conducted on different multimodal dataset, covering six types of modalities. Based on the results in Table 3, we further verify that the conventional Pareto method is inferior to the uniform baseline and loses its efficacy in multimodal scenarios. In addition, former

Table 4. Comparison of multi-task methods on NYUv2 dataset. Conv Pareto is the conventional Pareto method.

Method	Segmentation - mIoU (Higher Better)	Segmentation - Pix Acc (Higher Better)	Depth - Abs Err (Lower Better)	Depth - Rel Err (Lower Better)
Uniform baseline	25.80	52.68	0.6309	0.2680
Conv Pareto (Sener & Koltun, 2018)	**27.66**	53.91	0.6284	0.2685
GradNorm (Chen et al., 2018)	26.17	53.41	0.6219	0.2738
PCGrad (Yu et al., 2020)	26.44	53.93	0.6337	0.2658
MetaBalance (He et al., 2022)	27.04	**53.99**	0.6258	0.2677
MMPareto	26.35	53.48	**0.6216**	**0.2656**

Table 5. Results on MultiMNIST with 50% salt-and-pepper noise on the right part of images. Uniform baseline is a basic way where all losses are equally summed. Conv Pareto is the conventional Pareto method.

Method	Accuracy	
	Task 1	Task2
Uniform baseline	86.63	78.42
Conv Pareto (Sener & Koltun, 2018)	86.95	77.04 ($\downarrow$)
GradNorm (Chen et al., 2018)	85.95($\downarrow$)	79.52
PCGrad (Yu et al., 2020)	84.74($\downarrow$)	76.44($\downarrow$)
MetaBalance (He et al., 2022)	87.07	79.53
MMPareto	**87.72**	**80.64**

multi-task methods are also possibly invalid in the context of multimodal learning. For example, GradNorm method is inferior to the uniform baseline on both ModelNet40 and Kinetics Sounds dataset. In contrast, our MMPareto method, which specifically considers the multimodal learning properties, maintains its superior performance across various dataset with different kinds of modalities.

4.6. Extension to Multi-task Scenario

To evaluate scalability of our method in multi-task cases with similar property that there is a clear discrepancy in task difficulty, we conduct experiments on MultiMNIST dataset (Sabour et al., 2017). In MultiMNIST, two images with different digits from the original MNIST dataset are picked and then combined into a new one by putting one digit on the left and the other one on the right. Two tasks are to classify these two digits. To increase the difference in difficulty of tasks, we add 50% salt-and-pepper noise on the right part of images. Several data samples are provided in Appendix C. Based on Table 5, most multi-task methods have a performance drop (*e.g.,* GradNorm, PCGrad and conventional Pareto). Not surprisingly, our MMPareto could extend to this scenario and achieve considerable performance, indicating its ideal scalability.

Besides the MultiMNIST dataset, we also conduct experiments on the typical multi-task dataset, NYUv2. Widely-used benchmark (Liu et al., 2019) is used. We consider two tasks, semantic segmentation and depth estimation. As shown in Table 4, although our method is built on multimodal properties, it also achieves improvement on typical multi-task dataset, compared with uniform baseline. It even achieves the best results on depth estimation task. These results show the versatility of our method.

5. Discussion

In this paper, we identify previously ignored gradient conflicts in multimodal scenarios with discrepancies in learning difficulty of uni- and multimodal objectives, then propose MMPareto algorithm to diminish these conflict and alleviate imbalanced multimodal learning. Besides typical discriminative multimodal scenarios, more multimodal cases, like multimodal pre-training, have witnessed rapid development recently (Wang et al., 2023; Zhang et al., 2023; Chen et al., 2023). Our method is also expected to apply to these more general multimodal scenarios, like BLIP-2, easing the potential conflicts among multiple learning objectives.

Future work and limitation: We strengthen SGD noise term to improve model generalization, by adjusting gradient magnitude in this paper. In fact, the gradient covariance $\frac{1}{|S|}C$ of multiple losses and their correlation are also expected to provide a reliable reference when ensuring model generalization during gradient integration. Hence how to efficiently estimate and utilize these covariance is a promising next direction. Besides, our current convergence analysis is with relatively ideal assumptions. But in practice, it can greatly improve multimodal learning with a modest number of iterations in an acceptable range. This also inspires us to explore multimodal Pareto methods with more rigorous convergence theory.

Acknowledgements

Impact Statement

Our method could also contribute to our society in several aspects, *e.g.,* helping autonomous vehicle by enhancing multi-sensory learning.

References

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Chen, F.-L., Zhang, D.-Z., Han, M.-L., Chen, X.-Y., Shi, J., Xu, S., and Xu, B. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Désidéri, J.-A. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., and Zhao, H. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656. PMLR, 2023.

Fan, C., Chen, W., Tian, J., Li, Y., He, H., and Jin, Y. Maxgnr: A dynamic weight strategy via maximizing gradient-to-noise ratio for multi-task learning. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1178–1193, 2022.

Fan, Y., Xu, W., Wang, H., Wang, J., and Guo, S. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, Y., Feng, X., Cheng, C., Ji, G., Guo, Y., and Caverlee, J. Metabalance: improving multi-task recommendations via adapting gradient magnitudes of auxiliary tasks. In *Proceedings of the ACM Web Conference 2022*, pp. 2205–2215, 2022.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pp. 9226–9259. PMLR, 2022.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9012–9020, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, H., Li, X., Hu, P., Lei, Y., Li, C., and Zhou, Y. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22214–22224, 2023.

Lin, X., Zhen, H.-L., Li, Z., Zhang, Q.-F., and Kwong, S. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

Lin, Y.-M., Gao, Y., Gong, M.-G., Zhang, S.-J., Zhang, Y.-Q., and Li, Z.-Y. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, 20(4):539–553, 2023.

Liu, S., Johns, E., and Davison, A. J. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1871–1880, 2019.

Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Ma, P., Du, T., and Matusik, W. Efficient continuous pareto exploration in multi-task learning. In *International Conference on Machine Learning*, pp. 6522–6531. PMLR, 2020.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34: 14200–14213, 2021.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.

Trick, S., Koert, D., Peters, J., and Rothkopf, C. A. Multi-modal uncertainty reduction for intention recognition in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7009–7016. IEEE, 2019.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12695–12705, 2020.

Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., Tian, Y., and Gao, W. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023.

Wei, Y., Hu, D., Tian, Y., and Li, X. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.

Wei, Y., Feng, R., Wang, Z., and Hu, D. Enhancing multi-modal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Xu, R., Feng, R., Zhang, S.-X., and Hu, D. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yang, Z., Wei, Y., Liang, C., and Hu, D. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*, 2024.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836, 2020.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Zhang, L., Ruan, L., Hu, A., and Jin, Q. Multimodal pre-training from monolingual to multilingual. *Machine Intelligence Research*, 20(2):220–232, 2023.

Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

A. Parameter Description

For convenience, we include a table of parameter descriptions in Table 6.

B. Dataset and Experiment Settings

CREMA-D (Cao et al., 2014) is an audio-visual dataset for emotion recognition, including 7,442 video clips, each spanning 2 to 3 seconds in duration. The video content is that actors speak several short words. This dataset covers 6 emotions: angry, happy, sad, neutral, discarding, disgust and fear.

Kinetics Sounds (Arandjelovic & Zisserman, 2017) is an audio-visual dataset containing 31 human action classes selected from Kinetics dataset (Kay et al., 2017). All videos are manually annotated for human action using Mechanical Turk and cropped to 10 seconds long around the action.

Colored-and-gray-MNIST (Kim et al., 2019) is a synthetic dataset based on MNIST (LeCun et al., 1998). Each instance contains two kinds of images, a gray-scale and a monochromatic colored image. Monochromatic images in the training set are strongly color-correlated with their digit labels, while monochromatic images in the other sets are weakly color-correlated with their labels.

ModelNet40 (Wu et al., 2015) is a dataset with 3D objects, covering 40 categories. It contains 9,483 training samples and 2,468 test samples. This dataset could be used to classify these 3D objects based on the 2D views of their front-view and back-view data (Su et al., 2015). Data of all views is a collection of 2D images of a 3D object.

When not specified, ResNet-18 (He et al., 2016) is used as the backbone in experiments and models are trained from scratch. Concretely, for the visual encoder, we take multiple frames as the input, and feed them into the 2D network; for the audio encoder, we modified the input channel of ResNet-18 from three to one like (Chen et al., 2020) does and the rest parts remain unchanged; Encoders of other modalities are not modified. For the CNN backbone, we use the widely used late fusion method, to integrate unimodal features. For the Transformer backbone, MBT (Nagrani et al., 2021), is used as the backbone. Concretely, the backbone contains 6 single-modal layers and 2 layers with cross-modal interaction. Specifically, for the Colored-and-gray MNIST dataset, we build a neural network with 4 convolution layers and 1 average pool layer as the encoder, like (Fan et al., 2023) does. During the training, we use SGD with momentum (0.9) and set the learning rate at $1e - 3$. All models are trained on 2 NVIDIA RTX 3090 (Ti).

C. Samples of MultiMNIST Dataset

Here we provide several samples of MultiMNIST dataset. In MultiMNIST, two images with different digits from the original MNIST dataset are picked, and then combined into a new one by putting one digit on the left and the other one on the right. Two tasks are to classify these two digits. In order to increase the difference in difficulty between tasks, we add 50% salt-and-pepper noise on the right part of images.

(a) Sample 1 without noise; Label: [5,1].

(b) Sample 1 with noise; Label: [5,1].

(c) Sample 2 without noise; Label: [0,3].

(d) Sample 2 with noise; Label: [0,3].

Figure 6. Samples in MultiMNIST dataset.

D. More-than-two Modality Case

Our MMPareto method has no restriction on the number of modalities. To verify the method performance in more-than-two modality case, we conduct experiments on CMU-MOSI dataset (Zadeh et al., 2016). CMU-MOSI is a sentiment analysis dataset with three modalities, audio, video and text. It is annotated with utterance-level sentiment labels. This dataset consists of 93 movie review videos segmented into 2,199 utterances. According to results in Table 7, many existing methods for imbalanced multimodal learning problem only consider two modality case, and even can not extend to more modalities cases. Our MMPareto method is free of the limitation of the number of modalities and maintains effectiveness.

E. Proof for the Convergence of MMPareto

Remark 2. *The proposed MMPareto method admits an iteration sequence that converges to a Pareto stationarity.*

Table 6. The description of used parameters.

Parameter	Description
n	total number of modalities
$\mathcal{L}_m$	multimodal joint loss
$\mathcal{L}_u^k$	unimodal loss of $k-$th modality
N	number of training samples
X_i	input of $i-$th data sample
Y_i	label of $i-$th data sample
(X_i, Y_i)	$i-$th data sample
$\theta^k(t)$	parameters of unimodal encoder of the $k-$th modality at $t-$th iteration
S	set of $t-$th mini-batch
$\frac{1}{\|S\|}C^m$	batch sampling covariance of multimodal joint loss
$\frac{1}{\|S\|}C^u$	batch sampling covariance of unimodal loss
$\mathbf{g}_S^m$ / $\mathbf{g}_S^m(\theta^k(t))$	multimodal gradient of $t-$th iteration for encoder of $k-$th modality
$\mathbf{g}_S^u$ / $\mathbf{g}_S^u(\theta^k(t))$	unimodal gradient of $t-$th iteration for encoder of $k-$th modality
α^m	weight of multimodal gradient $\mathbf{g}_S^m$ in Pareto optimization problem
α^u	weight of unimodal gradient $\mathbf{g}_S^u$ in Pareto optimization problem
β	the angle between $\mathbf{g}_S^m$ and $\mathbf{g}_S^u$
η	learning rate of SGD optimization
γ	rescale factor of the integrated gradient of MMPareto method
$\mathbf{h}_S(\theta^k(t))$	the final integrated gradient of uniform baseline at $t-$th iteration
ϵ_t	SGD noise term of uniform baseline at $t-$th iteration
$\mathbf{h}_S^{\text{Pareto}}(\theta^k(t))$	the final integrated gradient of conventional Pareto method at $t-$th iteration
ζ_t	SGD noise term of conventional Pareto method at $t-$th iteration
$\mathbf{h}_S^{\text{MMPareto}}(\theta^k(t))$	the final integrated gradient of proposed MMPareto method at $t-$th iteration
ξ_t	SGD noise term of proposed MMPareto method at $t-$th iteration

Table 7. Comparison of imbalanced multimodal learning methods on CMU-MOSI dataset with three modalities. Greedy method could not extend to more-than-two modality case. * indicates that the original methods of OGM and PMR also only consider two modality cases, and we extend them while retaining core strategy.

Method	**Acc**	**macro F1**
One joint loss	75.07	73.16
Uniform baseline	75.95	73.93
G-Blending (Wang et al., 2020)	76.16	74.65
OGM* (Peng et al., 2022)	75.80	74.71
Greedy (Wu et al., 2022)	/	/
PMR* (Fan et al., 2023)	76.28	75.06
AGM (Li et al., 2023)	76.08	74.98
MMPareto	**76.53**	**75.59**

Proof. In MMPareto algorithm, at each training iteration, we first sovle the optimization problem:

$$\min_{\alpha^m, \alpha^u \in \mathcal{R}} \|\alpha^m \mathbf{g}_S^m + \alpha^u \mathbf{g}_S^u\|^2 \qquad (17)$$
$$s.t. \quad \alpha^m, \alpha^u \geq 0, \alpha^m + \alpha^u = 1.$$

For brevity, here we use $\{\mathbf{g}_S^i\}_{i \in \{m,u\}}$ to substitute $\{\mathbf{g}_S^i(\theta^k(t))\}_{i \in \{m,u\}}$. $\| \cdot \|$ denotes the L_2-norm. This problem is equal to finding the minimum-norm in the convex hull of the family of gradient vectors $\{\mathbf{g}_S^i\}_{i \in \{m,u\}}$. We denote the found minimum-norm as $\omega = \alpha^m \mathbf{g}_S^m + \alpha^u \mathbf{g}_S^u$. Based on (Désidéri, 2012), either ω to this optimization problem is 0 and the corresponding parameters are Pareto-stationary which is a necessary condition for Pareto-optimality, or ω can provide a descent directions common to all learning objectives. When the minimum-norm ω does not satisfy the condition of Pareto stationarity, we consider the non-conflict case and conflict case respectively.

We first analyze the non-conflict case, where $\cos \beta \geq 0$. β is the angle between $\mathbf{g}_S^m$ and $\mathbf{g}_S^u$. Under this case, we assign $2\alpha^m = 2\alpha^u = 1$. Then, the final gradient $\mathbf{h}_S^{\text{Pareto}} = \mathbf{g}_S^m + \mathbf{g}_S^u$ is with direction that can benefit all losses and enhanced generalization. Then we analyze the conflict case, where $\cos \beta < 0$. The results of optimization problem Equation 17 are used as α^m and α^u. Based on the above statement, we can have that $\mathbf{h}_S^{\text{Pareto}} = 2\alpha^m \mathbf{g}_S^m + 2\alpha^u \mathbf{g}_S^u$ can provide a direction that is common to all learning objective. Furthermore, we enhance the magnitude of final gradient to enhance SGD noise strength for improving model generalization. *In summary, the final gradient of MMPareto could always provide the direction that is common to all learning*

objectives.

If the training iteration stops in a finite number of steps, a Pareto-stationary point has been reached. Otherwise, the iteration continues indefinitely, generating an infinite sequence of shared parameters θ^k. *Since the value of loss function $\mathcal{L}_m$ and $\mathcal{L}_u^k$ is positive and monotone-decreasing during optimization, it is bounded. Hence, the sequence of parameter θ^k is itself bounded and it admits a subsequence converging to θ^{k^*}.*

Necessarily, θ^{k^*} is a Pareto-stationary point. In other words, the minimum-norm ω^* is zero at this step. To establish this, assume instead that the obtained minimum-norm ω^*, which corresponds to θ^{k^*}, is nonzero. A new iteration would potentially diminish each learning objective of a finite amount, and a better solution of parameter θ^k be found.

Overall, the proposed MMPareto method admits an iteration sequence that converges to a Pareto stationarity.

$\square$

F. Computation Cost and Convergence of MMPareto

In this section, we observe the time cost of gradient weight computation process and the convergence of our MMPareto method. In Table 8, we record the mean and variance of time cost per iteration. Based on the results, this process would not take a lot of time, without much effect on total training time.

In addition, we observe the number of iterations it takes to find the Pareto stationarity. According to results in Table 9, our method typically converges in a modest number of iterations in an acceptable range.

Table 8. Time cost of MMPareto calculation per iteration.

Dataset	Time cost mean	Time cost variance
CREMA-D	0.11s	1e-5
Kinetics Sounds	0.16s	1e-5
CG-MNIST	0.05s	1e-5
ModelNet40	0.10s	1e-5

Table 9. Number of iterations that find Pareto stationarity.

Dataset	Audio encoder	Video encoder
CREMA-D	#3098	#1709
Kinetics Sounds	#10197	#7957