G^{2}D: Boosting Multimodal Learning with Gradient-Guided Distillation

Mohammed Rakib, Arunkumar Bagavathi
Oklahoma State University
Stillwater, Oklahoma, United States
{mohammed.rakib, abagava}@okstate.edu

Abstract

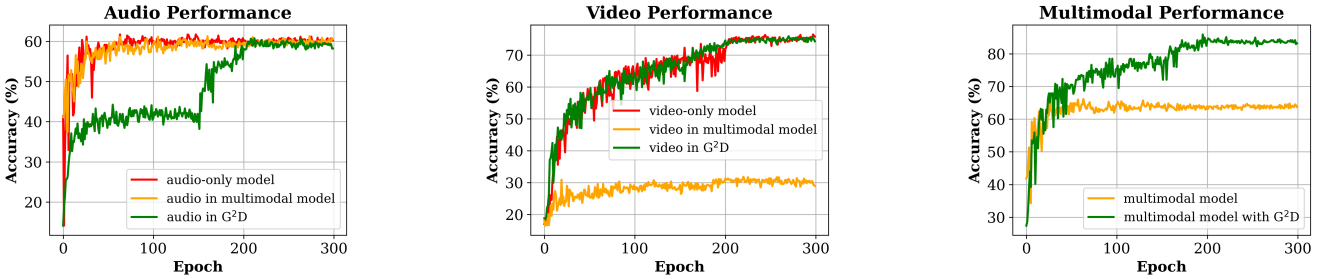
*Multimodal learning aims to leverage information from diverse data modalities to achieve more comprehensive performance. However, conventional multimodal models often suffer from modality imbalance, where one or a few modalities dominate model optimization, leading to suboptimal feature representation and underutilization of weak modalities. To address this challenge, we introduce Gradient-Guided Distillation (*G^{2}D*), a knowledge distillation framework that optimizes the multimodal model with a custom-built loss function that fuses both unimodal and multimodal objectives. G^{2}D further incorporates a dynamic sequential modality prioritization (SMP) technique in the learning process to ensure each modality leads the learning process, avoiding the pitfall of stronger modalities overshadowing weaker ones. We validate G^{2}D on multiple real-world datasets and show that G^{2}D amplifies the significance of weak modalities while training and outperforms state-of-the-art methods in classification and regression tasks. Our code is available here.*

1. Introduction

Multimodal learning is one of the most prominent multidisciplinary research areas due to the increasing demand to develop intelligent agents that perceive information from diverse sensory modalities. One of the primary challenges of multimodal learning models is the *modality imbalance* phenomenon [21, 24, 36, 41], also known as the modality competition [17, 18, 21] or modality laziness [8, 45]. Modality imbalance occurs when one modality dominates and other modalities are underutilized in the optimization of multimodal learning models. This causes (i) inferior multimodal performance compared to unimodal models [24, 29, 35], or (ii) a larger gap in individual modality when they are optimized jointly but still improve the model performance [9, 21]. This imbalance occurs due to poor alignment of the modalities, model overfitting to the modalities [35], and differences in the rate of model convergence [47]. An example of a modality imbalance with the multimodal dataset *CREMA-D* [4] is given in Figure 1. It is evident that the *audio* features(Figure 1a) dominate the video features(Figure 1b) when optimized in a multimodal fashion. However, video features give better performance with unimodal training. This results in sub-par performance with the joint multimodal training, shown in Figure 1c. In this work, our goal is to (i) increase the performance of weak modalities in multimodal settings and (ii) increase the overall performance of the multimodal model in downstream supervised tasks.

In recent years, many methods have been proposed to address modality imbalance in multimodal learning [8–10, 17, 21, 24, 35, 36, 38, 41, 42, 45]. *Gradient modulation* is one of the popular approaches in state-of-the-art methods to dynamically modify multimodal optimization gradients and maximize equal contributions from all modalities. A common form of gradient modulation is dynamically increasing the gradients of weak modalities only for late fusion [24, 42, 44] or for any type of fusion methods [9, 21] during the training process. Multiple variations of these gradient modulations also exist, for example, alternating gradients of each modality [45] and controlling the dominant modality gradients [10]. Although there exist multiple gradient modulation methods, there are very few methods that optimize both unimodal and multimodal learning objectives [8, 36] to add the benefits of both worlds. We aim to follow this trend in this work and introduce a novel optimization strategy by incorporating knowledge distillation.

In this work, we aim to utilize the full potential of both unimodal and multimodal learning for the given multimodal downstream task. We propose a framework *Gradient-Guided Distillation* (*G^{2}D*) that transfers knowledge from multiple unimodal teachers to multimodal student models. ***Novelty*** of *G^{2}D* is its use of knowledge distillation with a new learning objective and gradient modulation technique to mitigate modality imbalance and produce state-of-the-art results in multimodal learning. As depicted in Figure 1, G^{2}D improves the feature quality of multimodal encoder, allowing both the audio and video encoders to approach



(a) Audio Performance (b) Video Performance (c) Multimodal Performance

Figure 1. Performance of unimodal-only, unimodal in multimodal training, and purely multimodal models on the CREMA-D test set for multimodal classification. **(a)** audio modality is indifferent to training configurations; **(b)** video modality is vulnerable to the audio modality in a multimodal setting; **(c)** performance of the multimodal model is not optimal because of modality imbalance. G^{2}D limits the optimization of superior modality and enhances the video modality to optimize the multimodal performance.

the accuracy of their unimodal counterparts when integrated into the multimodal model. This leads to an overall more balanced and better-performing multimodal model while addressing the modality imbalance issue (Figure 1c). Our *three-fold* contributions in this work are:

- We introduce a knowledge distillation framework called G^{2}D that adapts a new optimization technique to fuse both unimodal and multimodal learning to enhance the performance of downstream tasks.
- We propose a new Sequential Modality Prioritization strategy that dynamically balances the optimization of weak and dominant modalities to mitigate imbalance.
- With extensive experiments on *six* real-world datasets, we show that G^{2}D minimizes modality imbalance and achieves superior performance. We also show that our approach is adaptive to existing methods.

2. Related Work

Knowledge distillation (KD) transfers knowledge from a larger and more complex model (*teacher*) to a smaller and more efficient model (*student*) [2, 16, 30]. Multimodal KD extends this concept by leveraging information from multiple modalities to enhance learning, which supports several real-world AI applications ([22, 26, 34, 39, 43]). Multimodal KD distills knowledge from one modality to improve performance of other modalities or in multimodal downstream tasks. Multimodal KD has demonstrated substantial benefits, including improved performance, better multimodal alignment, and enhanced generalization across modalities ([11, 19, 40, 43]). Multimodal KD has been used in multiple real-world problems like medical imaging [34] to address missing modality and action recognition [26] to transfer knowledge from multimodal ensemble to a unimodal model. In this paper, we build upon recent advances in multimodal KD [8], specifically targeting the modality imbalance problem during multimodal training. By utilizing unimodal teacher models to guide the multimodal student model, we ensure balanced learning across modalities.

Several methods have been proposed to mitigate the modality imbalance through gradient modulation, feature rebalancing, or modality-specific learning rate adjustments ([9, 10, 21, 24, 42, 44]). *Gradient modulation* techniques dynamically adjust gradients to balance modality contributions during training [21, 24]. However, these methods often require careful tuning of hyperparameters, which can limit their generalizability. *Feature rebalancing* aims to optimize multimodal interaction by adjusting the contribution of each modality by enhancing the performance of unimodal learners [17, 36, 37, 45]. There is another perspective of alleviating modality imbalance with a proper label fitting using contrastive learning [41]. In the context of multimodal KD, limited work has used unimodal teachers to supervise a multimodal student. UMT [7] directly supervises with unimodal teachers, while UME [8] aggregates their logits. Choosing between them requires empirical tuning, limiting adaptability across tasks and datasets. In this work, we propose the G^{2}D framework to address these limitations by combining multimodal knowledge distillation with a new gradient modulation technique. Unlike previous approaches that require careful hyperparameter tuning ([7, 8, 21, 24]), our approach dynamically suppresses dominant modalities based on insights from unimodal teachers. This not only makes our method applicable across different settings but also allows underrepresented modalities to learn more effectively.

3. Methodology

We propose *Gradient-Guided Distillation* (G^{2}D) to mitigate modality imbalance in multimodal learning through a combination of knowledge distillation and sequential modality prioritization. The proposed G^{2}D adapts to labeled multimodal datasets $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$,where each sample x_i consists of k-modality inputs, denoted as $x_i = (x_i^{m_1}, x_i^{m_2}, \ldots, x_i^{m_k})$, and an associated label y_i.

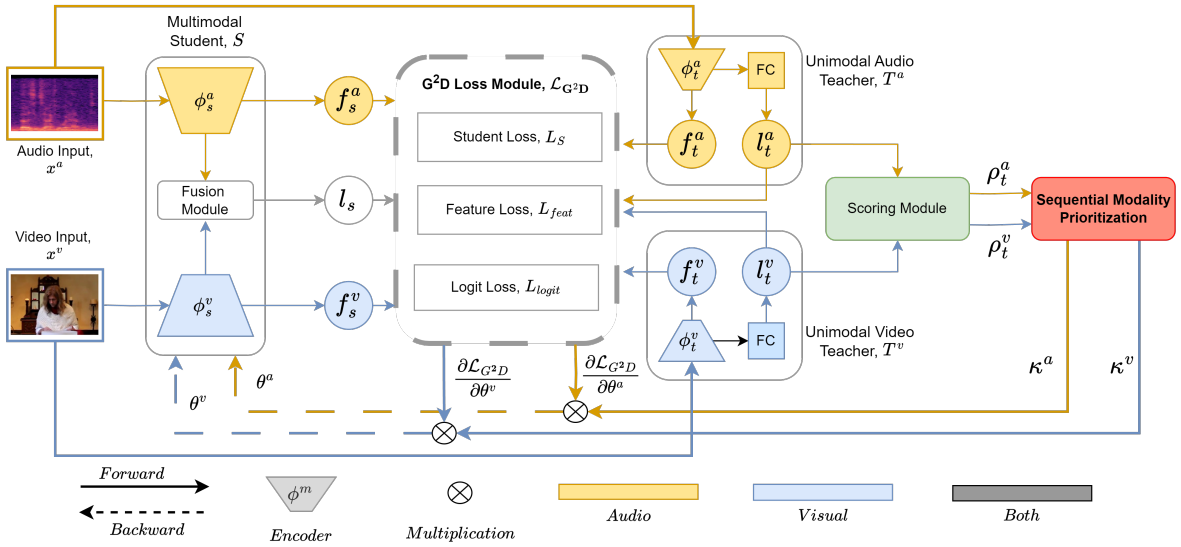


Figure 2. **G²D** consists of multiple, independently optimized *unimodal teacher* encoders and jointly optimized *multimodal student* encoders with all encoders generating feature representations and logits for each modality. The $\mathcal{L}_{\text{G}^2\text{D}}$ **Loss Module** consists of student loss, feature distillation loss, and logit distillation loss. Confidence scores from the **Scoring Module** are used by the **Sequential Modality Prioritization Module** to generate dynamic modulation coefficients that adaptively adjust the gradients of each encoder to ensure balanced contributions.

G²D, as illustrated in Figure 2, combines unimodal and multimodal learning by distilling the knowledge from unimodal teachers to the multimodal student with a new learning objective $\mathcal{L}_{\text{G}^2\text{D}}$. The *scoring module* in G²D determines modality-specific scores based on the knowledge of multiple unimodal teachers. The proposed *Sequential Modality Prioritization* dynamically identifies inferior modalities and empirically modulates the gradients of multimodal student encoders to mitigate modality imbalance.

3.1. G²D Loss Function

We extend traditional multimodal knowledge distillation [8] by introducing a new training objective ($\mathcal{L}_{\text{G}^2\text{D}}$) that combines *unimodal feature distillation* ($\mathcal{L}_{feat}$) loss and *unimodal logit distillation* ($\mathcal{L}_{\text{logit}}$) loss with the *multimodal student loss* ($\mathcal{L}_S$) to mitigate modality imbalance. We define the unimodal teacher model $\{T^m\}_{m=1}^k$ and the multimodal student model $[S]_1^k$. Each teacher model T^m is responsible for a single modality m and consists of an encoder ϕ_t^m parameterized by θ_t^m, which produces corresponding feature representations $f_t^m = \phi_t^m(x_i^m; \theta_t^m)$. We represent logits l_t^m of teacher models after passing f_t^m through a linear classifier ψ_t^m. Similarly, the student model $[S]_1^k$ processes the multimodal input x_i through modality-specific encoders ϕ_s^m, parameterized by θ_s^m, to obtain student feature representations $f_s^m = \phi_s^m(x_i^m; \theta_s^m)$. These features are then combined through a traditional multimodal fusion module $\Phi_{\text{fusion}}(f_s^{m_1}, f_s^{m_2}, \ldots, f_s^{m_k})$, to produce a multimodal representation, which is used to calculate multimodal student logits $l_s = \psi_s(\Phi_{\text{fusion}}(f_s^{m_1}, f_s^{m_2}, \ldots, f_s^{m_k}))$. The proposed $\mathcal{L}_{\text{G}^2\text{D}}$ comprises three key components:

1. **Multimodal Student Loss ($\mathcal{L}_S$)** is a supervised loss to map multimodal inputs x_i to the label y_i:

$$\mathcal{L}_S = \mathbb{E}_{(x,y)\sim[\mathcal{D}]_{m=1}^k}\left[\ell(p, y)\right] \tag{1}$$

where ℓ is the cross-entropy loss ($\ell(p, y) = -\sum_{w=1}^{C} y_w log(p)$) for C-class classification tasks with $p = l_s(x; \theta_s)$ or mean squared error ($\ell(p, y) = \frac{1}{N}\sum_{i=1}^{N}(p - y)^2$) for regression tasks with $p = \sigma(l_s(x; \theta_s))$, where σ is the sigmoid function.

2. **Feature Distillation Loss** ($\mathcal{L}_{\text{feat}}$): To prevent the student model from discarding information from weaker modalities, we impose an L2-based feature alignment loss that minimizes the discrepancy between multimodal student and unimodal teachers' feature representations:

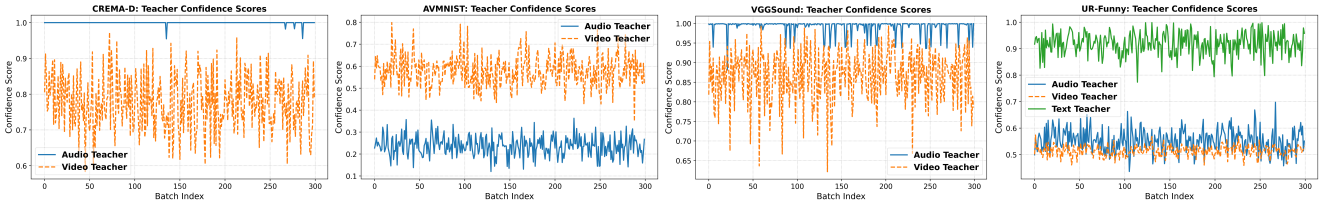
$$\mathcal{L}_{\text{feat}}^m = \mathbb{E}_{x\sim\mathcal{D}}\left[\|\phi_s^m(x^m; \theta_s^m) - \phi_t^m(x^m; \theta_t^m)\|^2\right] \tag{2}$$

where ϕ_s^m and ϕ_t^m are student and teacher features, respectively, as functions of input x^m and encoder parameters.

3. **Logits Distillation Loss ($\mathcal{L}_{\text{logit}}$)**: Logit-based distillation enables the student model to capture class-level relationships and decision boundaries defined by the teacher. We integrate a new logit-based distillation using Kullback-Leibler (KL) divergence [16] that learns the distribution from unimodal teachers to the multimodal student:

$$\mathcal{L}_{\text{logit}}^m = \mathbb{E}_{x\sim\mathcal{D}}\left[\text{KL}\left(\sigma(l_t^m(x^m; \theta_t^m))\|\sigma(l_s(x; \theta_s))\right)\right] \tag{3}$$

where σ denotes the softmax function.



(a) CREMA-D (b) AV-MNIST (c) VGGSound (d) UR-Funny

Figure 3. **Unimodal teacher confidence scores across multimodal datasets.** Each line is the confidence of a specific modality (audio, visual, or text). Modality bias on all datasets, with higher scores for one modality, motivates our use of sequential modality prioritization.

We define the G^{2}D loss as:

$$\mathcal{L}_{\text{G}^2\text{D}} = \mathcal{L}_S + \alpha \sum_{m=1}^{k} \mathcal{L}^m_{\text{feat}} + \beta \sum_{m=1}^{k} \mathcal{L}^m_{\text{logit}} \tag{4}$$

where α and β are weighting coefficients for the feature loss and the logit loss, respectively. This formulation enables the multimodal student model to leverage the strengths of each unimodal teacher model effectively. Feature distillation loss ensures that the student retains modality-specific representations, while logit distillation loss aligns the student's predictions with teacher distributions, capturing inter-class and intra-class dependencies. Although this new learning objective enables maximization of quality, multimodal models can still be biased to dominant modalities as they optimize all modalities simultaneously. We introduce an adaptive training strategy based on unimodal confidence scores to mitigate modality imbalance.

3.2. Quantifying Modality Confidence

Scoring mechanisms are widely used to measure the contributions of individual modalities in multimodal learning [21, 24]. As unimodal models are outside the bounds of modality imbalance, we utilize their confidence in determining the imbalance ratio. Our scoring module in G^{2}D quantify the confidence ρ of each unimodal teacher T^m as the batch-wise average of their softmax function:

$$\rho^m_t = \frac{1}{|\mathcal{B}^m|} \sum_{(x^m_i, y^m_i) \in \mathcal{B}^m} \text{Softmax}(l^m_t(x^m_i; \theta^m))[y^m_i], \tag{5}$$

where $|\mathcal{B}^m|$ represents the number of m modality data samples in the batch, and $\text{Softmax}(l^m_t(x^m_i; \theta^m))[y^m_i]$ is the probability assigned to the ground truth label y^m_i by the teacher model T^m for the sample x_i.

The score ρ^m_t serves as an indicator of how confident the unimodal teacher T^m is in predicting the correct label for the given batch. A higher score indicates greater confidence, signifying that modality m dominates in multimodal training. This information is then used to guide gradient updates in the student model, dynamically adjusting training to mitigate the dominance of any single modality.

3.3. Modulating Gradients with Sequential Modality Prioritization (SMP)

Multimodal datasets often undergo modality overfitting during training [35] and give priority only to dominant modalities, limiting the optimization of weaker modalities. Using the modality scores ρ^m, we find this trend on all four classification datasets, as shown in Figure 3. It is evident that one modality consistently exhibits higher confidence scores, indicating a modality imbalance and dominance over other modalities in the given downstream task. We also note that this pattern is not grounded to a specific modality. For example, in CREMA-D (Figure 3a), the audio teacher has higher scores than the video teacher, and in UR-Funny (Figure 3d), the text modality remains dominant over audio and visual inputs. These findings align with previous research that identifies modality bias as a prevalent issue in multimodal datasets [1, 13, 46]. This further leads to modality imbalance in multimodal learning, as analyzed in Figure 1. To address this issue, we hypothesize the following:

Hypothesis 1. *Leveraging the confidence scores of unimodal models to determine less confident modalities and sequentially prioritizing them during multimodal training can mitigate modality imbalance.*

To test this hypothesis, we propose *sequential modality prioritization* strategy for the multimodal training in our proposed G^{2}D framework. For each training iteration q, we rank confidence scores of all teacher models ρ^m_t. If this ranked list is given as π_t, then $\pi_t[1]$ corresponds to the least confident modality, and $\pi_t[k]$ corresponds to the most dominant modality. We aim to generate an automatic training schedule process to identify the set of prioritized modalities $\mathcal{M}_q$ based on the modality confidence scores π_t, as given in Equation (6).

$$\mathcal{M}_q = \begin{cases} \{\pi_t[1]\} & \text{for } 1 \le e \le \tau_1 \\ \{\pi_t[2]\} & \text{for } \tau_1 < e \le \tau_1 + \tau_2 \\ \vdots & \\ \{\pi_t[k-1]\} & \text{for } \sum_{j=1}^{k-2} \tau_j < e \le \sum_{j=1}^{k-1} \tau_j \\ \{\pi_t[1], \ldots, \pi_t[k]\} & \text{for } \sum_{j=1}^{k-1} \tau_j < e \le \sum_{j=1}^{k} \tau_j \end{cases} \tag{6}$$

where e is the training epoch, and τ_j is the hyperparameter set to denote the number of epochs for optimizing j-th prioritized modality, where j is the index of the ranked list π_t. The modulation coefficients κ_q^m for each modality m identifies the modality to optimize in multimodal learning, as given in Equation (7).

$$\kappa_q^m = \begin{cases} 1 & \text{if modality } m \in \mathcal{M}_q, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

The gradient update for the parameters θ_q^m of modality m for iteration q in the multimodal student model S is then:

$$\theta_{q+1}^m = \theta_q^m - \eta \cdot \kappa_q^m \cdot \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}} \left[\frac{\partial \mathcal{L}_{\text{G}^2\text{D}}(x_i, y_i)}{\partial \theta_q^m} \right] \tag{8}$$

where η is the learning rate, and $\mathcal{L}_{\text{G}^2\text{D}}$ represents the total loss function as defined in Equation (4).

During each epoch range τ_j, only the corresponding modality $\pi_t(j)$ is assigned $\kappa_q^m = 1$, while all others are set to 0, ensuring that the prioritized modality receives full gradient updates. After the prioritized phases for all less confident modalities, the most confident modality ($\pi_t(k)$) is trained jointly with all other modalities, with $\kappa_q^m = 1$ for all m. This sequential prioritization strategy allows each modality to lead the learning process in turn, thereby mitigating persistent modality dominance. We summarize the complete multimodal training with G^{2}D in Algorithm 1.

Algorithm 1 Multimodal Learning with G^{2}D

Input: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, unimodal teachers $\{T^m\}_{m=1}^k$, multimodal student S, iterations q
Initialize student model S with random weights
Load pre-trained weights for teacher models $\{T^m\}_{m=1}^k$
for iteration $i = 0, \ldots, q-1$ **do**
  Sample a fresh mini-batch (x, y) from $\mathcal{D}$
  Feed-forward the batch through S to obtain $\{f_s^m\}_{m=1}^k$ and l_s
  Feed-forward each modality of a batch through $\{T^m\}_{m=1}^k$ to get $\{f_t^m\}_{m=1}^k$ and $\{l_t^m\}_{m=1}^k$
  Compute $\mathcal{L}_{\text{G}^2\text{D}}$ loss using Eq. (4)
  Calculate ρ_t^m for each modality using Eq. (5)
  Calculate κ^m using Eq. (7)
  Find modality gradients $\frac{\partial \mathcal{L}_{\text{G}^2\text{D}}}{\partial \theta^m}$ and update parameters θ^m for each modality m using Eq. (8)
end for
return Trained multimodal student model S

4. Experiments

We evaluate the G^{2}D framework on the basis of the following questions: **Q1:** How does G^{2}D compare to state-of-the-art methods for addressing modality imbalance and overall multimodal performance in supervised tasks? **Q2:** How does G^{2}D influence the modality gap in multimodal learning?, **Q3:** Does G^{2}D enhance feature space alignment between unimodal and multimodal models?, **Q4:** How SMP is influencing the multimodal learning?, and **Q5:** Which fusion and suppression techniques are the best for G^{2}D?

4.1. G^{2}D Evaluation

4.1.1. Experimental Setup

Datasets. We chose five multimodal classification datasets and one regression dataset. **CREMA-D** [4] is an audio-visual dataset for speech emotion recognition with six emotion classes. **AV-MNIST** [33] is a synthetic dataset with PCA-projected MNIST images and audio spectrograms for ten-digit classes. **VGGSound** [6] is a large-scale audio-visual dataset with 309 classes of everyday audio events, featuring video clips of 10 seconds each. **UR-Funny** [14] is a binary classification dataset for humor detection that incorporates text, visual gestures, and acoustic modalities. **IEMOCAP** [3] is an audio-visual-text dataset for emotion recognition in dyadic conversations. **MIS-ME** [27] is a regression dataset containing raw soil patch images and corresponding meteorological data for soil moisture estimation. To the best of our knowledge, we are the *first* to evaluate the modality imbalance in a multimodal regression setting and use tabular data for multimodal learning problems.

Baselines. We compare G^{2}D with ten state-of-the-art methods that address modality imbalance, including MSES [10], MSLR [42], AGM [21], PMR [9], OGM-GE [24], MLA [45], MM-Pareto [36], ReconBoost[17], DLMG[41], and UMT [8]. We compare the baseline methods across four datasets that include audio, visual, and text modalities. For the regression task, we compare our approach to MISME [27], which estimates soil moisture from soil patch images and meteorological data. We evaluate the performance of each baseline in both unimodal and multimodal contexts for each modality. Please refer to **??** for more detailed baseline descriptions.

Backbone and Hyperparameter Settings. For audio-visual datasets (CREMA-D, VGGSound, and AV-MNIST), we use ResNet-18 [15] as the encoder for both audio and video modalities in the teacher and student models. For the UR-Funny and IEMOCAP dataset, which involves audio, visual, and text modalities, we use a Transformer-based encoder [32] for each modality. For MIS-ME, we adopt MobileNetV2 [28] as the image feature extractor and use the fully connected neural network proposed in [27] for processing tabular meteorological data. To ensure a fair comparison, we use identical backbone architectures across all baseline models and employ late fusion for training. All models are optimized using SGD with a batch size of 16 and trained on a single NVIDIA A10 GPU. More details on experimental settings are provided in **??**.

Table 1. Performance comparison on various audio-visual datasets reported in accuracy (%). "Multi" represents the evaluation of the multimodal student model, while "Audio" and "Video" rows indicate the performance of modality-specific encoders within the multimodal model. T_a and T_v denote unimodal evaluations for the audio and video teacher models, respectively.

Dataset		T^a	T^v	Joint-Train	MSES	MSLR	AGM	PMR	OGM-GE	MLA	MM Pareto	Recon Boost	DLMG	UMT	G^{2}D (Ours)
CREMA-D	Audio	61.69	-	59.95	54.86	54.86	48.58	49.19	58.60	59.27	65.46	57.71	54.37	61.02	56.45
	Video	-	76.48	27.42	22.57	26.31	57.85	23.25	49.06	64.91	55.24	65.21	70.89	25.40	72.72
	Multi	-	-	67.47	60.99	64.42	78.48	59.13	72.18	79.70	75.13	79.82	<u>83.62</u>	67.61	**85.89**
AV-MNIST	Audio	42.69	-	16.05	27.50	22.72	38.90	37.60	24.53	42.26	42.11	41.50	41.99	31.55	39.10
	Video	-	65.44	55.83	63.34	62.92	63.65	58.50	55.85	65.30	65.26	64.28	65.04	64.08	65.09
	Multi	-	-	69.77	70.68	70.62	72.14	71.82	71.08	65.32	<u>72.63</u>	72.47	72.14	72.33	**73.03**
VGGSound	Audio	43.39	-	39.22	39.57	39.10	38.15	26.30	37.96	37.56	42.44	42.35	41.54	42.12	39.43
	Video	-	32.32	18.70	17.85	18.66	25.65	7.12	22.64	32.02	17.94	18.12	23.65	23.77	29.88
	Multi	-	-	50.97	50.76	50.98	47.11	33.07	51.45	51.65	49.69	50.97	52.74	<u>53.78</u>	**53.82**

4.1.2. Results

In this section, we present the accuracy(%) of all models with best results in **bold** and second best results <u>underlined</u>.

G^{2}D on Two Modalities and Audio-Visual Domain. Table 1 presents the following key observations:

1. Unimodal performances (T_a and T_v) and joint-training reveal that modality imbalance is dataset-dependent. On CREMA-D and VGGSound, video performs well in the unimodal setup but becomes suppressed in multimodal training, while it is the opposite in AV-MNIST as audio is underutilized in multimodal setup. This confirms that modality imbalance is a prevalent issue in multimodal learning, leading to suboptimal fusion in joint-training.

2. DLMG, ReconBoost, and gradient modulation methods (AGM, OGM-GE, MLA, and MMPareto) attempt to reduce modality imbalance and improve fusion. While effective, they do not fully bridge the discrepancy among modalities, as imbalance persists across datasets. G^{2}D surpasses all baselines, demonstrating that SMP ensures balanced optimization and better multimodal integration.

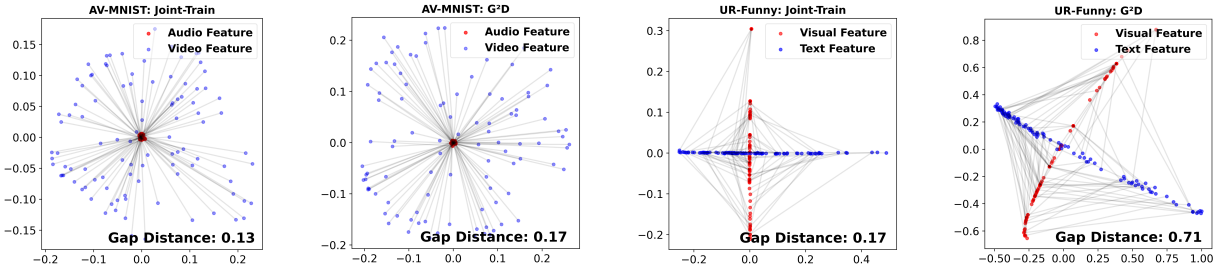
3. To the best of our knowledge, UMT is the only knowledge distillation-based baseline addressing modality imbalance. The results show that the proposed G^2D loss that distills knowledge from unimodal teachers with the dynamic training strategy using SMP gives better optimization for weak modalities to outperform UMT across all datasets.

G^{2}D on Three Modalities and Text Domain. Unlike prior approaches [9, 38], G^{2}D is not constrained by the number of modalities. We now analyze results with a three-modality dataset on UR-FUNNY, given in Table 2. In this special case of experiment, we compare our results with baseline models using all combinations of modalities. We find with joint-training that text modality is dominant in all multimodal settings and incorporating all modalities leads to the best multimodal performance. (i) **G^{2}D performance:** We observe that G^{2}D consistently outperforms methods incorporating adaptive training strategies, such as OGM-GE, MMPareto, and ReconBoost, as well as the KD-based approach UMT, demonstrating its effectiveness in mitigating modality imbalance. (ii) **Modality depression:** Another interesting finding with more than two modalities (**A-V-TXT**) is the dominant modality (text) depression across most of our baseline models. We suspect that these models give over-prioritization to weak modalities while not allowing the required optimization for dominant modalities. G^{2}D, on the other hand, treats all modalities fairly with the proposed SMP to reduce modality depression.

Table 2. Accuracy (%) on the UR-Funny dataset across all modality combinations. Unimodal teacher performance for audio, visual, and text are 61.57%, 58.25%, and 61.77%, respectively.

Type		Joint-Train	OGM-GE	MM Pareto	Recon Boost	UMT	G^{2}D (Ours)
A-V	Audio	57.34	59.76	61.77	60.53	54.63	59.05
	Visual	53.92	53.82	55.73	57.87	56.44	58.05
	Multi	61.57	61.87	61.27	<u>62.07</u>	60.46	**62.98**
A-TXT	Audio	50.30	54.12	58.15	50.18	55.63	59.86
	Text	57.44	58.35	58.45	56.98	57.75	58.85
	Multi	62.17	62.47	<u>62.88</u>	61.06	62.47	**63.28**
V-TXT	Visual	49.30	55.33	56.04	55.41	56.34	56.34
	Text	51.21	58.95	59.15	50.94	53.82	56.04
	Multi	62.07	62.98	61.27	60.07	<u>63.18</u>	**63.48**
A-V-TXT	Audio	55.03	50.30	58.05	51.65	50.70	59.15
	Visual	54.93	55.73	56.14	55.26	54.93	55.94
	Text	58.25	55.71	58.55	56.25	52.72	58.15
	Multi	62.58	<u>63.68</u>	62.88	61.37	63.38	**65.49**

G^{2}D on the Multimodal Regression Task. To evaluate the robustness and real-world applicability of G^{2}D, we applied it to a soil moisture estimation task using an in-wild raw soil patch dataset [27], captured via cameras. As given in Table 3, we find that modality imbalance occurs in regression tasks as well. We find that G^{2}D outperforms the baseline method MIS-ME, indicating its effectiveness in multimodal regression tasks.



(a) AV-MNIST: Joint-Train (b) AV-MNIST: G^{2}D (c) UR-Funny: Joint-Train (d) UR-Funny: G^{2}D

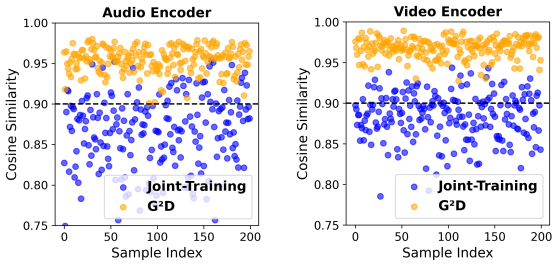
Figure 4. Modality gap for AV-MNIST and UR-Funny datasets, with G^{2}D increasing the modality separation compared to joint-training.

Table 3. G^{2}D for Soil Moisture Regression Task on MIS-ME Dataset with *tabular* and *image* Modalities

Metrics	Tabular	Image	Joint-Train	MIS-ME	G^{2}D
MAPE	15.49	8.22	14.62	7.52	**7.01**
R^2	0.34	0.76	0.42	0.80	**0.82**

4.2. Analysis of G^{2}D

Analyzing Modality Gap. Prior work has shown that multimodal learning leads to distinct modality-specific embeddings, with a larger *modality gap* often correlating with improved performance [23, 31]. Following these insights, we visualize the modality gap in AV-MNIST (audio-visual) and UR-Funny (visual-text) datasets, as shown in Figure 4. Compared to joint-training (Figures 4a and 4c), G^{2}D (Figures 4b and 4d) results in a more pronounced modality gap, making modalities more distinguishable in the embedding space. This separation is particularly evident in text-inclusive models, facilitating better feature utilization. These results highlight G^{2}D's ability to mitigate modality bias and enhance multimodal learning by preserving modality-specific characteristics.



(a) Audio Encoder (b) Video Encoder

Figure 5. Alignment between unimodal and multimodal features in the audio encoder (Figure 5a) and the video encoder (Figure 5b).

Analyzing Feature Alignment of G^{2}D. We first analyze the robustness of multimodal features by aligning them with unimodal teacher features for both audio and video encoders in CREMA-D. In this experiment, we use cosine similarity to measure the feature alignment. With Figure 5a and Figure 5b, we show that the alignment of both modalities between unimodal and multimodal features is consistently higher for the G^{2}D compared to a simple joint-training without KD. We consider that better feature alignment in G^{2}D is one of the important factors in improving modality imbalance in multimodal learning.

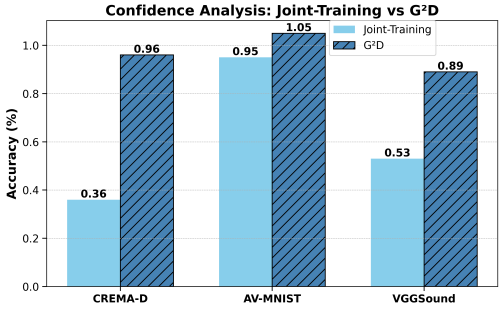


Figure 6. Analyzing Confidence Ratio of Joint-Training with G^{2}D

Analyzing Modality Imbalance with Confidence Ratio. To quantify the impact of modality imbalance, we compute *confidence ratio* to measure the weaker modality's confidence relative to its unimodal teacher. Specifically, we first compute the average confidence score ρ for the weaker modality across the entire dataset in both joint-training and G^{2}D. Then, we normalize these values by dividing them by the corresponding unimodal teacher's confidence score. A lower confidence ratio indicates that the weaker modality is overshadowed during training, while a higher ratio suggests that the weaker modality in multimodal setting is reaching its performance close to that of its unimodal setup. As shown in Fig. 6, G^{2}D consistently yields higher confidence ratios than joint-training, demonstrating its ability to mitigate modality discrepancy by ensuring a more balanced optimization process and preventing weaker modalities from being suppressed in the multimodal training.

Further analysis of G^{2}D, including distinction from key baselines and computational cost, is in **??**.

4.3. Ablation Study

Table 4. Effect of SMP on Different Multimodal Methods

Method	SMP	CREMA-D	AV-MNIST	UR-Funny
Joint-Train	✗	67.47	69.77	62.58
	✓	80.78	72.51	63.58
UMT	✗	67.61	72.33	63.38
	✓	82.39	72.68	64.59
G^{2}D loss	✗	78.63	72.76	63.78
	✓	**85.89**	**73.03**	**65.49**

Impact of SMP. The results in Table 4 demonstrate *two-pronged* observations on SMP with three datasets. (i) SMP integration enhances the performance of both vanilla joint-training and the multimodal KD-based approach UMT. (ii) Incorporating SMP to G^{2}D loss achieves the best overall performance, outperforming all baselines, highlighting its role in mitigating modality imbalance and improving multimodal learning. This confirms the effectiveness and adaptability of SMP, not just in G^{2}D but also in other models.

Table 5. Performance comparison of G^{2}D using different fusion strategies across multiple datasets in terms of accuracy (%).

Fusion	CREMA-D	AV-MNIST	VGGSound	UR-Funny
Sum	81.59	72.70	50.67	63.08
Concat	83.60	72.98	53.40	64.49
FiLM [25]	84.27	72.73	48.11	63.48
BiGated [20]	81.32	72.89	46.66	63.38
Cross-Attention [5]	85.35	72.96	53.58	65.09
Late Fusion [12]	**85.89**	**73.03**	**53.82**	**65.49**

G^{2}D with Various Fusion Modules. Table 5 compares the effect of different fusion strategies on G^{2}D. Out of the traditional fusion methods used for the modality imbalance problem, *late fusion* consistently achieves the best results. This highlights the effectiveness of leveraging independent unimodal representations while preserving their distinct contributions. However, *Cross-Attention* based fusion closely follows the performance of late fusion, demonstrating its ability to enhance cross-modal interactions by dynamically attending to relevant features. Concat fusion provides strong performance but falls slightly behind Late Fusion. FiLM and BiGated offer adaptive feature integration yet fail to match the top-performing methods. Descriptions of these fusion strategies are in **??**.

Modality Suppression in G^{2}D. We compare two suppression strategies: partial suppression and complete suppression. Partial suppression follows OGM-GE [24], applying gradient modulation with $1 - \tanh(x)$, where x is the ratio of modality scores. Complete suppression utilizes SMP, which zeroes out gradients of dominant modalities, allowing the suppressed modality to train to convergence. Table 6 shows that complete suppression consistently outperforms partial suppression across all datasets. This improvement arises from SMP's ability to shift optimization focus toward the suppressed modality, preventing interference and enabling more effective multimodal feature integration. So, complete suppression significantly mitigates modality imbalance and enhances multimodal learning.

Table 6. Partial vs. Complete Modality Suppression in G^{2}D

Type	CREMA-D	AV-MNIST	VGGSound	UR-Funny
Partial	81.99	72.83	51.16	63.68
Complete	**85.89**	**73.03**	**53.82**	**65.49**

Table 7. Effect of τ_j on G^{2}D with two & three modalities

(τ_1, τ_2)	**(0,150)**	**(50,150)**	**(100,150)**	**(150,150)**
CREMA-D	78.63	82.80	83.74	**85.89**

(τ_1, τ_2, τ_3)	**(0,0,150)**	**(50,50,150)**	**(75,75,150)**
IEMOCAP	75.30	76.99	**77.19**

Effect of Prioritization Epochs (τ_j). We analyze the impact of τ_j, the hyperparameter controlling the epochs for each SMP stage, in Table 7. For instance, with two modalities ($k = 2$), the schedule (τ_1, τ_2) denotes epochs for training the weakest modality alone, followed by a joint training phase for both. Similarly, for three modalities ($k = 3$), (τ_1, τ_2, τ_3) defines epochs for the weakest alone, then the second weakest alone, and finally a joint phase for all three. The results in Table 7 show that systematically increasing the dedicated epochs for weaker modalities consistently improves G^{2}D's performance. This finding strongly validates our Hypothesis 1 (Sec. 3.3) that providing weaker modalities with dedicated, interference-free training phases is crucial for mitigating modality imbalance.

Additional ablations are provided in **??**.

5. Conclusion

In this paper, we presented G^{2}D, a simple but effective novel framework designed to tackle modality imbalance in multimodal learning through gradient-guided distillation and sequential modality prioritization. By fusing unimodal and multimodal learning objectives together with knowledge distillation and utilizing confidence scores from unimodal teachers, G^{2}D dynamically prioritizes weaker modalities. With these contributions, G^{2}D ensures each modality contributes effectively during training without being overshadowed by any dominant modalities. Our experimental results, which span multiple classification datasets and a regression task, illustrate that G^{2}D enhances feature alignment, mitigates modality imbalance, and outperforms existing state-of-the-art methods. We believe that the proposed gradient modulation strategy holds great potential to advance balanced learning in complex multimodal scenarios, paving the way for more inclusive and robust AI systems.

References

[1] Ibrahim M. Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiao-Qi Zhai. Clip the bias: How useful is balancing data in multimodal learning? *ArXiv*, abs/2403.04547, 2024. 4

[2] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Knowledge Discovery and Data Mining*, 2006. 2

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. 5

[4] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. 1, 5

[5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356, 2021. 8

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020. 5

[7] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers, 2021. 2

[8] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8632–8656. PMLR, 2023. 1, 2, 3, 5

[9] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20029–20038, 2023. 1, 2, 5, 6

[10] Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-specific learning rate control for multimodal classification. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pages 412–422. Springer, 2020. 1, 2, 5

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. In *International Journal of Computer Vision*, pages 1789–1819. Springer, 2021. 2

[12] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, pages 3437–3443 Vol. 4, 2005. 8

[13] Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan S. Kankanhalli, and A. Bimbo. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19:1 – 22, 2022. 4

[14] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China, 2019. Association for Computational Linguistics. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3

[17] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *The Forty-first International Conference on Machine Learning*, 2024. 1, 2, 5

[18] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022. 1

[19] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[20] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 8

[21] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22214–22224, 2023. 1, 2, 4, 5

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, pages 9694–9705, 2020. 2

[23] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 7

[24] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022. 1, 2, 4, 5, 8

[25] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 8

[26] Gorjan Radevski, Zhipeng Luo, Yifei Zhu, Luc Van Gool, and Dengxin Dai. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12123–12133, 2023. 2

[27] Mohammed Rakib, Adil Aman Mohammed, D. Cole Diggins, Sumit Sharma, Jeff Michael Sadler, Tyson Ochsner, and Arun Bagavathi. Mis-me: A multi-modal framework for soil moisture estimation, 2024. 5, 6

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE CVPR*, pages 4510–4520, 2018. 5

[29] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021. 1

[30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. 2

[31] Vishaal Udandarao. *Understanding and Fixing the Modality Gap in Vision-Language Models*. Phd thesis, University of Cambridge, 2022. 7

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5

[33] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: A multilayer approach for multimodal fusion. In *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, page 575–589, Berlin, Heidelberg, 2019. Springer-Verlag. 5

[34] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part IV*, page 216–226, Berlin, Heidelberg, 2023. Springer-Verlag. 2

[35] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 1, 4

[36] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance, 2024. 1, 2, 5

[37] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV*, page 71–86, Berlin, Heidelberg, 2024. Springer-Verlag. 2

[38] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022. 1, 6

[39] Hongwei Xu, Goutham Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Multimodal-cop: Self-supervised vision-language pre-training with auxiliary tasks. *arXiv preprint arXiv:2107.07773*, 2021. 2

[40] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding cross-modal knowledge distillation. In *International Conference on Learning Representations*, 2023. 2

[41] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, pages 62108–62122. Curran Associates, Inc., 2024. 1, 2, 5

[42] Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834, 2022. 1, 2, 5

[43] Yuan Yuan et al. Multiscale knowledge distillation with attention based fusion for robust human activity recognition. *Scientific Reports*, 14:63195, 2024. 2

[44] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1, 2

[45] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27456–27466, 2024. 1, 2, 5

[46] Yedi Zhang, Peter E. Latham, and Andrew M. Saxe. A theory of unimodal bias in multimodal learning. *ArXiv*, abs/2312.00935, 2023. 4

[47] Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, and Ken Zheng. Balancing multimodal learning via online logit modulation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5753–5761, 2024. 1