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

Supplementary Material

A. Appendix

A.1. Detailed Dataset Description

A.1.1. Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [4]

CREMA-D is a multimodal dataset designed for emotion recognition research. It contains audio-visual recordings of actors portraying a variety of emotional states, including Anger, Disgust, Fear, Happy, Neutral, and Sad. The dataset features actors from diverse racial and ethnic backgrounds, covering a wide age range, which makes it suitable for studying the interplay between audio and visual emotional expressions. Ratings for emotional intensity and accuracy were gathered from crowd-sourced participants. The dataset is divided into a training set of 6,027 samples, a validation set of 669 samples, and a test set of 745 samples, facilitating robust model training and evaluation.

A.1.2. Audio Visual MNIST (AV-MNIST) [37]

AV-MNIST is a synthetic multimodal dataset designed for audio-visual digit classification. It combines visual MNIST digit images, downsampled using PCA to retain 25% of their original energy, with audio samples of spoken digits from the TIDigits dataset [22]. The audio samples are represented as 112 × 112 spectrograms and are augmented with noise from the ESC-50 dataset [28]. The dataset consists of 70,000 audio-visual pairs, including 55,000 for training, 10,000 for testing, and 5,000 selected from the training set for validation.

A.1.3. VGGSound [6]

VGGSound is a large-scale audio-visual dataset designed for training and evaluating audio recognition models. It consists of over 200,000 video clips sourced from YouTube, each containing audio-visual correspondence where the sound source is visually present in the video. The dataset includes 310 diverse classes covering various real-world environments, such as people, animals, music, and nature. Each clip is 10 seconds long, ensuring both the audio and visual elements are aligned, making it ideal for audio-visual learning tasks.

A.1.4. UR-Funny [15]

UR-FUNNY is a multimodal dataset created for the task of humor detection, utilizing text, visual gestures, and prosodic acoustic cues. The dataset comprises 1,866 video clips collected from TED Talks featuring diverse speakers and covering 417 different topics. Each clip is labeled with binary humor annotations, with an equal number of humorous and non-humorous samples, ensuring a balanced dataset. The multimodal nature of UR-FUNNY makes it particularly suitable for investigating the relationships among different modalities, offering insights into how text, vision, and audio can jointly contribute to understanding humor in a multimodal learning context.

A.1.5. IEMOCAP [3]

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is a widely used resource for emotion recognition, containing approximately 12 hours of audio-visual data from ten actors engaged in scripted and improvised dyadic conversations to elicit a range of authentic emotions. The dataset is rich in modalities, providing synchronized video, speech, facial motion capture data, and text transcriptions. Each utterance is annotated by multiple raters for both categorical emotions and dimensional attributes (valence, arousal, and dominance), making it suitable for diverse and nuanced modeling tasks. Due to its naturalistic dyadic interactions, IEMOCAP serves as a benchmark for developing emotion-aware conversational AI and understanding complex multimodal emotional cues.

A.2. Details on Baselines

A.2.1. Modality-Specific Early Stopping (MSES) [11]

MSES is a multimodal learning approach that aims to prevent overfitting by independently managing the learning rate for each modality. Within the MSES framework, each modality's classifier is treated as a separate task, and early stopping is employed when a modality begins to overfit, while others continue to learn. By formulating a multitask setup, MSES allows for the independent regulation of modality-specific learning progress, ensuring that the stronger modalities do not overshadow the weaker ones during joint training. This method effectively prevents overfitting by identifying and stopping learning for modalities when their validation loss fails to improve, thereby maximizing balanced contributions from each modality and enhancing the overall multimodal learning process.

A.2.2. Modality-Specific Learning Rate (MSLR) [46]

MSLR aims to optimize multimodal late-fusion models by assigning unique learning rates to each modality instead of using a single global learning rate. This approach helps prevent the vanishing gradients issue that can occur when learning rates are not tailored to the specific characteristics of each modality. By assigning modality-specific learning rates, MSLR ensures that each modality contributes effectively to the learning process, ultimately improving the overall performance of the multimodal model.

A.2.3. Adaptive Gradient Modulation (AGM) [23]

AGM addresses modality competition in multimodal models by modulating the participation level of each modality during training. Inspired by Shapley value-based attribution and the OGM-GE algorithm, AGM isolates the contribution of individual modalities and modulates their gradient update intensity accordingly, allowing stronger modalities to be suppressed while amplifying weaker ones. This adaptive strategy applies to all types of fusion architectures, thereby boosting the overall model performance by ensuring a balanced contribution from each modality and mitigating dominance effects that lead to suboptimal joint training outcomes.

A.2.4. Prototypical Modality Rebalance (PMR) [9]

PMR tackles the ”modality imbalance” issue by applying different learning strategies to each modality to ensure more balanced learning. Specifically, PMR uses prototypical cross-entropy (PCE) loss to accelerate the slow-learning modality, allowing it to align more closely with prototypical representations, while also reducing the inhibition from dominant modalities via prototypical entropy regularization (PER). The method effectively exploits features of each modality independently and helps prevent one modality from dominating the learning process, thereby enhancing overall multimodal learning performance.

A.2.5. On-the-fly Gradient Modulation with Generalization Enhancement (OGM-GE) [26]

OGM-GE addresses the issue of under-optimization for specific modalities in multimodal learning by dynamically modulating gradient contributions for each modality. This approach balances the learning pace by modulating gradients of modality-specific coefficients during backpropagation, reducing the dominance of stronger modalities and facilitating better feature exploitation of weaker ones. Additionally, OGM-GE incorporates a generalization enhancement mechanism, adding dynamic Gaussian noise to improve model generalization.

A.2.6. Multimodal Learning with Alternating Unimodal Adaptation (MLA) [51]

MLA addresses the issue of modality dominance by alternating the training focus between modalities rather than using conventional joint optimization. This alternating unimodal adaptation helps avoid interference between modalities, allowing each to reach its full potential while still maintaining cross-modal interactions through a shared head. A gradient modification mechanism is introduced to mitigate ”modality forgetting,” thereby preserving cross-modal information learned during previous iterations. At inference, MLA integrates multimodal information dynamically, using uncertainty-based fusion to manage imbalance across modality-specific contributions effectively.

A.2.7. MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance [40]

MMPareto aims to enhance multimodal learning by addressing the gradient conflict that arises between unimodal and multimodal learning objectives. The algorithm uses Pareto integration to align gradient directions across objectives, ensuring a final gradient that benefits all modalities without compromising any. By balancing gradient direction and boosting gradient magnitude, MMPareto improves generalization, providing ”innocent unimodal assistance” to enhance the performance of each modality while maintaining the consistency of multimodal learning.

A.2.8. On Uni-Modal Feature Learning in Supervised Multi-Modal Learning (UMT) [8]

This paper addresses the problem of insufficient learning of unimodal features in multi-modal learning. The proposed framework consists of two approaches: Uni-Modal Teacher (UMT) and Uni-Modal Ensemble (UME). UMT distills unimodal pre-trained features into a multi-modal model during late-fusion training, ensuring that the representations learned for each modality are preserved effectively while maintaining cross-modal interactions. UME, on the other hand, avoids cross-modal interactions by combining the predictions from unimodal models directly, thus preventing negative interference. To decide which approach to use, they employ an empirical trick explained in the paper. In our experiments, we compare against UMT due to its use of knowledge distillation (KD), which aligns with our proposed approach.

A.2.9. ReconBoost: Boosting Can Achieve Modality Reconcilement [18]

ReconBoost introduces a modality-alternating learning paradigm to mitigate modality competition in multimodal learning. Instead of optimizing all modalities simultaneously, ReconBoost updates each modality separately, ensuring that weaker modalities are not overshadowed by stronger ones. A KL-divergence-based reconcilement regularization is incorporated to maximize diversity between current and past updates, aligning the method with gradient-boosting principles. Unlike traditional boosting, ReconBoost only retains the most recent learner per modality, preventing overfitting and excessive reliance on strong modalities. Additionally, it integrates a memory consolidation regularization to preserve historical modality-specific information and a global rectification scheme to refine joint optimization. Empirical results across multiple benchmarks demonstrate that ReconBoost effectively reconciles modality learning dynamics, leading to improved multimodal fusion performance.

A.2.10. Facilitating Multimodal Classification via Dynamically Learning Modality Gap (DLMG) [45]

DLMG addresses the modality imbalance problem in multimodal learning by focusing on disparities in category label fitting across different modalities. Unlike prior methods that primarily regulate learning rates or gradient contributions, DLMG leverages contrastive learning to align modality representations and reduce dominance effects. The approach dynamically integrates supervised classification loss and contrastive modality matching loss through either a heuristic strategy or a learning-based optimization strategy that adjusts their relative importance during training. By progressively refining modality alignment while maintaining label supervision, DLMG minimizes performance gaps between dominant and non-dominant modalities, leading to a more balanced and effective multimodal learning process.

A.2.11. Detached and Interactive Multimodal Learning (DI-MML) [10]

DI-MML proposes that modality competition is a direct result of the uniform learning objective used in traditional joint training frameworks. To eliminate this competition, DI-MML proposes a detached learning framework where each modality's encoder is trained separately with its own isolated learning objective. To enable cross-modal interaction without reintroducing competition, the framework employs two key strategies: (1) a shared classifier is used to align features from different modalities into a common embedding space, and (2) a novel Dimension-decoupled Unidirectional Contrastive (DUC) loss is introduced. The DUC loss identifies "effective" and "ineffective" feature dimensions within each modality and then transfers knowledge unidirectionally from the effective dimensions of one modality to the corresponding ineffective dimensions of another. This strategy facilitates the exchange of complementary information while preserving the integrity of each modality's well-learned features.

A.3. Details on Fusion Techniques

A.3.1. Summation

Summation fusion is a straightforward multimodal integration technique where features from multiple modalities are combined through element-wise addition. Each modality contributes independently, and their respective representations are directly summed without any complex cross-modal interactions. In this approach, the output of each modality-specific encoder is first processed by a fully connected layer to generate unimodal predictions, which are then added together to form a unified representation. This combined output is used to compute a loss, which subsequently updates all components involved, including the encoders and the fully connected layers. Summation fusion's strength lies in its simplicity and ease of implementation, as it does not require intricate fusion mechanisms. However, it does not explicitly capture inter-modal relationships, potentially limiting its effectiveness in scenarios where richer cross-modal interactions are beneficial.

A.3.2. Concatenation

Concatenation fusion is a common strategy for integrating information from different modalities by concatenating their feature vectors along a specified axis. This method combines feature representations directly, allowing the model to consider information from all modalities together as a single, extended vector. Despite enabling joint perception of multimodal data, it does not explicitly model cross-modal interactions. The concatenated features are passed through a fully connected layer, where the input size equals the sum of all encoder output dimensions, and the output size matches the number of classes. During training, the model uses the resulting fusion output to compute the loss and update all the involved parameters, including those of the individual encoders and the fully connected layer. Concatenation fusion is effective in creating a unified feature representation, but it relies on subsequent layers to extract and learn any interactions between the modalities.

A.3.3. Feature-wise Linear Modulation (FiLM) [27]

FiLM is a sophisticated fusion method that integrates information from multiple modalities by adjusting feature representations in one modality according to the information from another. This modulation approach uses conditional inputs to produce parameters that scale and shift feature activations, enabling the model to dynamically adjust its processing based on context. FiLM works by passing the conditioning modality through a layer that outputs these modulation parameters, which then directly adjust the target modality's features before they proceed to the next layers in the model. By providing targeted feature modulation, FiLM helps capture cross-modal nuances and allows the model to be more adaptive in multimodal learning tasks that require context-sensitive adjustments.

A.3.4. BiLinear Gated Fusion (BiGated) [21]

BiGated fusion combines bilinear pooling and gating mechanisms to enhance the integration of multiple modalities by capturing their complex interactions. This technique explicitly models cross-modal relationships, providing a more expressive and fine-grained fusion strategy compared to simpler approaches like concatenation or summation. In BiGated fusion, each modality passes through its own fully connected layer, much like summation. However, what sets BiGated apart is its use of a gating mechanism—one modality's hidden state is processed through an activation function (we use sigmoid) to generate a gated weight, which is then used to modulate the contributions of other modalities. This approach ensures that each modality can dynamically influence how other modalities are represented in the fusion process, allowing for a richer and more adaptive multimodal representation before proceeding to the final classification layers.

A.3.5. Cross-Attention Fusion [5]

Cross-attention fusion enables the dynamic and adaptive integration of multiple modalities by allowing each modality to attend to others through bidirectional or all-directional attention mechanisms. This approach explicitly models intermodal dependencies, ensuring that each modality can selectively focus on the most relevant features from others. In our implementation, for two-modal cases, modality X attends to modality Y and vice versa, refining their representations based on mutual interactions. For three-modal scenarios, all-directional attention is applied, where each modality interacts not only with one other but also with the third, ensuring comprehensive multimodal integration. The attended representations are normalized to enhance stability and mitigate potential imbalances in feature contributions. Finally, the refined features from all modalities are projected into a unified representation through a fully connected layer. This mechanism effectively captures nuanced cross-modal relationships, allowing the model to leverage complementary modality-specific information for robust multimodal learning.

A.3.6. Late Fusion [13]

Late fusion technique involves independently processing each modality through its respective model or encoder, followed by combining the outputs at a later stage to produce the final prediction. This approach allows each modality to be modeled and optimized in isolation, maintaining the unique properties of each data source. However, it may miss opportunities to exploit early cross-modal interactions that could provide additional benefits during feature learning. In late fusion, each modality-specific encoder is followed by its own fully connected layer, which is trained solely on that modality's data. The fusion output is computed by averaging the outputs of all unimodal models, ensuring that the fusion occurs only after independent learning is complete. This independence provides flexibility and robustness, especially in scenarios where some modalities may be missing, but limits the ability to deeply integrate multimodal relationships early in the learning process.

A.4. Details on Experimental Setups

A.4.1. Model Architectures

ResNet-18 ResNet-18, a convolutional neural network with 18 layers, belongs to the ResNet family and is renowned for addressing the vanishing gradient problem through residual connections. These residual connections allow information to bypass some layers, which helps stabilize training even in deeper networks. In our experiments, we use ResNet-18 as an encoder for both audio and video modalities across CREMA-D, AV-MNIST, and VGGSound datasets. We used a specific weight initialization strategy: Xavier normal for fully connected layers, Kaiming normal for convolutional layers, and constant initialization for batch normalization layers, which facilitated an effective starting point for network training and ensured stable convergence across multimodal tasks.

Transformer Transformers are powerful architectures designed for handling sequential data and capturing long-range dependencies through self-attention mechanisms. In our implementation, we employ Transformers as encoders for the audio, video, and text modalities of the UR-Funny and IEMOCAP dataset. Following the approach described in [23], we utilized a 4-layer Transformer encoder with eight attention heads and a hidden dimension of 768 for each modality in the UR-Funny and IEMOCAP dataset. Input features were projected to a 768-dimensional embedding using a convolutional layer, ensuring consistency across different modalities. We employed a similar initialization strategy to ResNet-18 to facilitate stable training.

A.4.2. Hyperparameters

We trained our models on 1 Nvidia A10 GPU with a batch size of 16 using the SGD optimizer, with a momentum of 0.9 and a weight decay of 1×10^{-4}. We initialized the learning rate at 0.001 and decayed it by a ratio of 0.1 every 200 epochs. For all experiments, we set the random seed to 999 for reproducibility. We defined the G^{2}D loss function as a weighted sum of student loss, feature loss, and logit loss, where α and β are weighting coefficients for the feature loss and logit loss, respectively. We set both α and β to 1.0 for all datasets. Additionally, for the logit loss, we used a temperature of 1.0 in the KL Divergence without further softening, effectively utilizing hard logits for the training process.

A.5. Comparison of G^{2}D with DI-MML

In response to reviewer feedback, we provide an additional comparison against the state-of-the-art baseline DI-MML [10]. As shown in Table 8, G^{2}D outperforms DI-MML on both the CREMA-D and AV-MNIST datasets. This result further validates the effectiveness of G^{2}D in relation to current leading methods in the field.

Table 8. Comparing G^{2}D with DI-MML

Method	Joint-Train	DI-MML	G^{2}D
CREMA-D	67.47	83.51	**85.89**
AV-MNIST	69.77	71.35	**73.03**

Table 9. Comparing Components of G^{2}D with UMT & OGM-GE

Method	Joint-Train	UMT	G^{2}D Loss	OGM-GE	SMP	G^{2}D (SMP + G^{2}D Loss)
CREMA-D	67.47	67.61	78.63	72.18	80.78	**85.89**
AV-MNIST	69.77	72.33	72.76	71.08	72.51	**73.03**

Table 10. Single-Batch Resource Metrics on CREMAD

Method	Total Memory (MB)	Total Execution Time (ms)
Joint-Train	12.1366 MB	4531.8
G^{2}D	12.1998 MB	4539.8

A.6. Further Analysis of G^{2}D

A.6.1. Distinction of G^{2}D from UMT and OGM-GE

The primary novelty of G^{2}D arises from its unique *G^{2}D loss* and its Sequential Modality Prioritization (SMP) technique, and critically, from their synergistic combination.

Our *G^{2}D loss* improves upon the distillation strategy of UMT [8] by incorporating a KL divergence-based logit loss. This addition is crucial for enabling the student model to learn the nuanced inter- and intra-class relationships captured by the unimodal teachers' soft logits.

Furthermore, our SMP technique is fundamentally different from the gradient modulation in OGM-GE [26] in two principal ways:

1. **Guidance for Modulation:** OGM-GE calculates modality confidence from the student's own encoders during training. This signal can be noisy and unreliable, especially in early stages. In contrast, SMP leverages stable confidence scores from *pre-trained unimodal teachers*, providing a more robust and accurate signal to identify weaker modalities automatically.
2. **Suppression Mechanism:** OGM-GE uses functions like $1 - \tanh(\cdot)$ to only *partially* suppress dominant modalities, meaning they continue to train simultaneously and modality competition can persist. SMP enforces a *complete gradient shutdown* for non-prioritized modalities. This ensures that the prioritized weak modality trains in true isolation, more effectively mitigating interference from dominant modalities.

The results in Table 9 validate these distinctions, showing that the *G^{2}D loss* alone surpasses UMT, SMP alone surpasses OGM-GE, and their combination yields the best overall performance.

A.6.2. Synergy of Distillation and Sequential Modality Prioritization

The motivation for integrating our *G^{2}D loss* (via KD) with SMP is to address the limitations of using either technique alone. The distillation component leverages unimodal teachers—trained in isolation—to provide the student with stable, competition-free feature and logit targets. This guides the student towards more balanced representations than learning solely from GT labels amidst modality competition.

However, even with this guidance, the student's modality encoders are still optimized simultaneously, which allows modality imbalance to persist. SMP is introduced to solve this. The crucial *synergy* lies in the fact that during the isolated training phases enforced by SMP, the prioritized weak modality learns not only from the ground-truth labels but also from the rich, interference-free knowledge distilled from its unimodal teacher via the *G^{2}D loss*. This focused, dual-signal learning in an isolated context enables the robust development of weaker modalities. By combining SMP with our distillation objective, G^{2}D mitigates modality imbalance more thoroughly and effectively than using either technique independently.

A.6.3. Computational Cost

The training overhead of G^{2}D is negligible, and there is no additional overhead during inference. This efficiency stems from the framework's design: unimodal teacher models are pre-trained, and their outputs (e.g., logits and features) are saved. During the multimodal student model's training, these pre-computed outputs are loaded from disk per batch in a process analogous to loading the dataset itself. As quantified in Table 10, for a typical 16-sample batch, G^{2}D requires only $\approx$0.5% more memory and adds merely $\approx$0.15% to the execution time compared to a standard joint-training baseline. Therefore, in resource-constrained environments, if a traditional joint-training model is feasible, G^{2}D is also readily viable by performing the one-time teacher training first and then training the student.

A.7. Additional Ablation Studies

A.7.1. Learning with Missing Modalities

To evaluate the robustness of G^{2}D with incomplete data, we conduct experiments on the IEMOCAP dataset with randomly missing modalities, generating miss rate masks from 0% to 60% following the setup in MLA [51]. As demonstrated in Table 11, G^{2}D consistently outperforms several state-of-the-art methods designed specifically for this task across all tested rates. This superior performance, even as data becomes highly sparse, suggests that by mitigating modality imbalance and fostering well-rounded representations, our framework learns more resilient features that are

Table 11. G^{2}D vs. Missing Modality Methods on IEMOCAP

Miss Rate	Joint-Train	CRA [34]	MMIN [53]	CPM-Net [49]	TATE [48]	G^{2}D
0%	75.51	76.21	74.94	58.00	69.92	**77.19**
20%	69.06	67.34	69.36	53.65	63.22	**71.49**
40%	61.09	57.04	63.30	51.01	60.36	**65.10**
60%	52.41	43.22	57.52	47.38	57.99	**61.50**

less dependent on any single data source, making it inherently more robust when modalities are unavailable.

Table 12. Effect of α and β in G^{2}D with SMP

Dataset	(α, β) **weights**						
	(0, 0)	(0.25, 0.75)	(0.5, 0.5)	(0.75, 0.25)	(1, 0)	(0, 1)	(1, 1)
CREMA-D	80.78	84.41	84.95	84.81	82.39	84.68	**85.89**
UR-Funny	63.58	64.79	64.29	64.29	64.59	64.69	**65.49**

A.7.2. Effect of α and β in G^{2}D.

α and β denote the weighting coefficients of feature loss and logit loss, respectively in the proposed G^{2}D loss. Table 12 evaluates the effect of changing the weightage of feature loss and logit loss on G^{2}D. Assigning full weight to both losses ($\alpha = 1, \beta = 1$) yields the best overall performance, highlighting their combined importance in multimodal learning. In contrast, removing both losses ($\alpha = 0, \beta = 0$) significantly reduces performance, confirming their necessity. While different weight combinations impact results, incorporating both losses with higher weight leads to greater improvements across datasets.