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

Accepted at the International Conference on Computer Vision (ICCV), 2025

Mohammed Rakib

mohammed.rakib@okstate.edu

PhD Student
Reasoning and Artificial Intelligence (rAIson) Lab
Department of Computer Science

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Outline

Presentation Roadmap

Background: What is Multimodal Learning?

Learning from Diverse Data Sources

Multimodal learning builds models that process and relate information from multiple *data types*, or *modalities*.

The goal is to create a more comprehensive understanding, much like how humans use sight, hearing, and touch together.

The standard pipeline involves three key steps:

- **Encoding:** Extract features from each input.
- **Fusion:** Combine the features into one.
- **Prediction:** Use the fused data for a task.

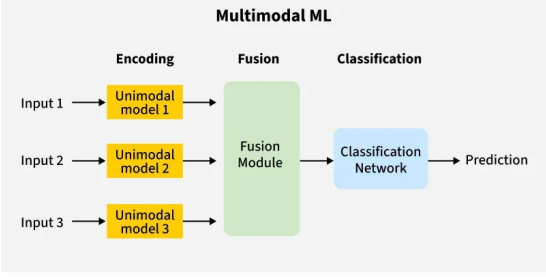


Figure: A standard multimodal learning pipeline.

Background: What is Knowledge Distillation?

Learning from an Expert Teacher

Knowledge Distillation (KD) is a technique where a *compact student model* learns from a *larger teacher model*, **transferring knowledge through logits, features, or intermediate representations.**

- The student mimics the teacher's outputs, learning not just the "what" (labels) but the "how" (richer patterns) *[Hinton et al., 2015]*.
- This process "distills" the teacher's generalized knowledge into the smaller student *[Gou et al., 2021]*.
- Initially used for **model compression**, KD is now vital for complex multimodal tasks like cross-modal knowledge transfer & handling missing data *[Wang et al., 2023]*.

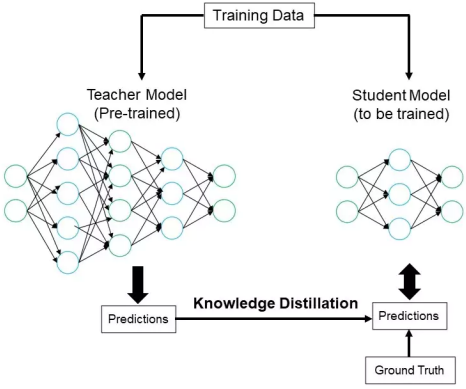


Figure: The Knowledge Distillation (KD) framework.

The Challenge: Modality Imbalance

When One Modality Dominates the Learning Process

The Core Problem:

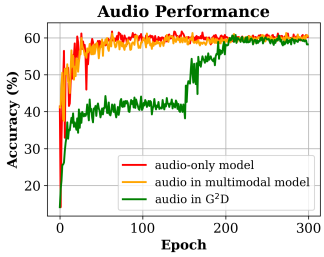
- In multimodal models, one modality often contains *stronger signals* or *learns much faster* than others.
- This phenomenon, known as **modality imbalance** or **modality competition**, causes the ”stronger” modality to dominate the joint training process *[Peng et al., 2022]*.

The Consequence:

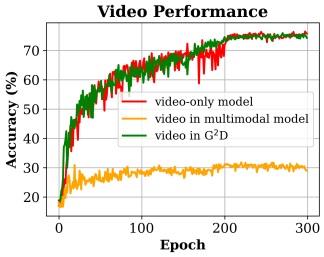
- Weaker modalities are **underutilized**, preventing the model from learning a truly robust, fused representation.
- Leads to **suboptimal performance** that can be even worse than using a single modality alone.

The Challenge: Modality Imbalance

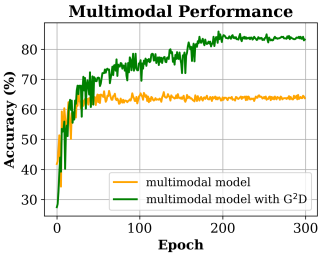
An Example on the CREMA-D Dataset



(a) Audio Performance



(b) Video Performance



(c) Multimodal Performance

Figure: In a standard model, the **strong** audio modality performs well, but the **weak** video modality is suppressed (yellow line, middle graph). This severely harms the final multimodal performance (yellow line, right graph).

Our Solution: Gradient-Guided Distillation (G^2D)

Actively Balancing Modalities to Boost Performance

ICCV OCT 19-23, 2025 HONOLULU HAWAII

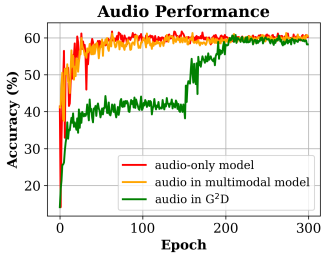
Our goal is to create a framework that mitigates this imbalance and actively boosts the performance of weaker modalities.

We introduce **Gradient-Guided Distillation (G^2D)**, which combines two key ideas:

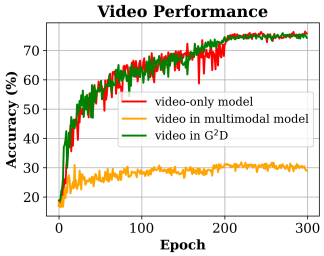
- **Knowledge Distillation:** We transfer knowledge from expert *unimodal teacher* models to a single *multimodal student* model.
- **Sequential Modality Prioritization (SMP):** We use a dynamic training strategy that gives each modality—especially the weaker ones—a dedicated turn to lead the learning process.

Our Solution: Gradient-Guided Distillation (G^2D)

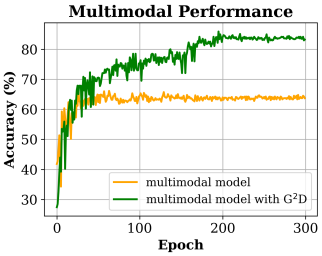
A Glimpse of the Results



(a) Audio Performance



(b) Video Performance



(c) Multimodal Performance

Figure: With G^2D, the weak video modality's performance is rescued and significantly improved (green line, middle graph). This leads to a substantial boost in the final multimodal accuracy (green line, right graph).

Related Work

Common Strategies for Mitigating Modality Imbalance

Most state-of-the-art methods address modality imbalance through two primary ways:

- **Gradient Modulation:** The most common approach. It dynamically adjusts modality gradients to suppress dominant inputs and amplify weaker ones.
  - Popular methods include OGM-GE *[Peng et al., 2022]*, AGM *[Li et al., 2023]*, and using modality-specific learning rates (MSLR) *[Yao and Mihalcea, 2022]*.
- **Feature Rebalancing & Alternating Training:** These methods optimize interactions by alternating the training focus between modalities (MLA *[Zhang et al., 2024]*) or using specialized losses to accelerate the learning of weaker modalities (PMR *[Fan et al., 2023]*).
- **Common Limitation:** Extensive hyperparameter tuning limiting their generalizability.

Our Contribution

G^2D combines KD with a novel gradient modulation technique called Sequential Modality Prioritization (SMP) that uses robust signals from unimodal teachers, removing the need for extensive manual tuning.

Methodology: Gradient-Guided Distillation (G^2D)

High-Level Architecture

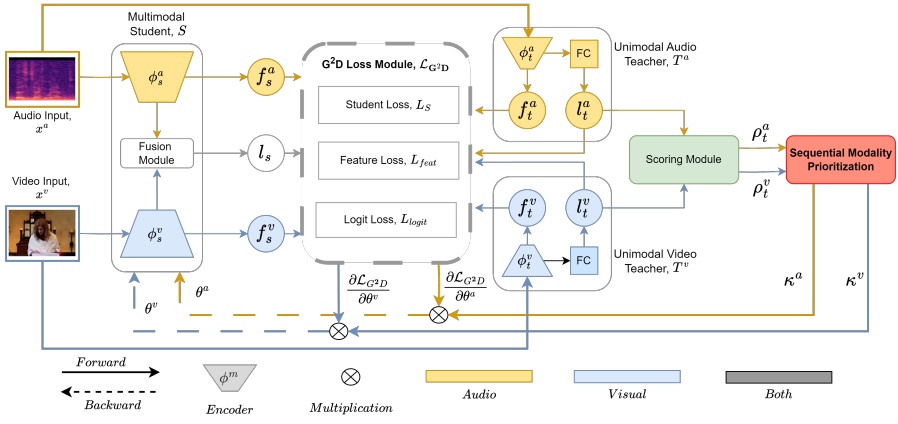


Figure: **Overview:** G^2D uses pre-trained unimodal teachers (right) to guide a multimodal student (left). Knowledge is transferred via our custom G^2D **Loss Module**. The **Scoring Module** calculates teacher confidence, which the **Sequential Modality Prioritization (SMP)** module uses to dynamically modulate the student's gradients, ensuring balanced learning.

Methodology: The G^2D Loss Function

Combining Three Key Objectives

Our total loss, $\mathcal{L}_{G^2D}$, combines a standard student loss with two distillation losses that leverage the unimodal teachers.

1. Student Loss ($\mathcal{L}_S$) Standard

supervised loss mapping the *student's* final multimodal prediction to the **ground-truth (GT)** label.

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

2. Feature Loss ($\mathcal{L}_{\text{feat}}$) L2-loss

aligns *student* and *teacher* **features** for each modality preventing weaker modalities from being ignored

$$\mathcal{L}^m_{\text{feat}} = \mathbb{E}_{x\sim\mathcal{D}}\left[\|\phi^m_s - \phi^m_t\|^2\right] \quad (2)$$

3. Logit Loss ($\mathcal{L}_{\text{logit}}$) KL-divergence

loss aligns the output distribution of *student* with each of the teacher's **logits** transferring class-relationships

$$\mathcal{L}^m_{\text{logit}} = \mathbb{E}_{x\sim\mathcal{D}}\left[\text{KL}(\sigma(l^m_t)\|\sigma(l_s))\right] \quad (3)$$

Total G²D Loss: The final loss is a weighted sum of the three components:

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

Methodology: Architecture Recap

Revisiting the G²D Framework

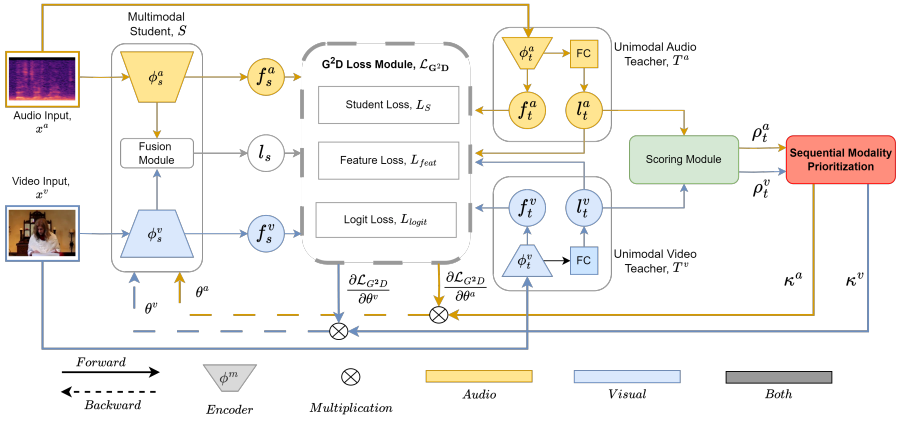
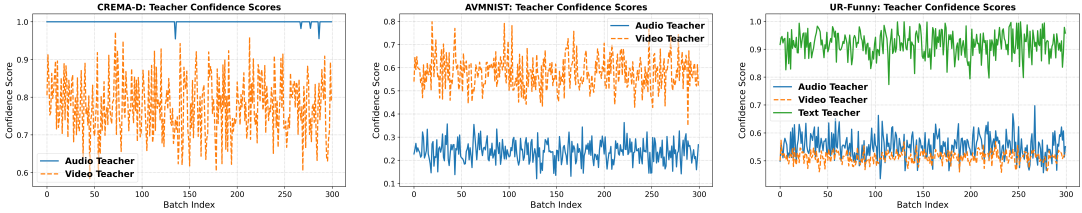


Figure: Having defined the G^2D **Loss Module**, we now focus on how the framework dynamically balances modalities, starting with the **Scoring Module**.

Methodology: Quantifying Modality Confidence

Using Unimodal Teachers as a Stable Signal



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

Figure: Teacher confidence scores on three datasets. The consistent gap between modalities demonstrates a clear bias, which motivates our prioritization strategy.

- We use the pre-trained **unimodal teachers** as a stable signal to determine which modality is dominant for a given batch of data.
- The confidence score ρ_t^m is the batch-wise average probability assigned to the ground-truth label:

$$\rho_t^m = \frac{1}{|\mathcal{B}^m|} \sum_{(x_i^m, y_i^m) \in \mathcal{B}^m} \text{Softmax}(l_t^m(x_i^m; \theta^m))[y_i^m] \tag{5}$$

Methodology: Architecture Recap

From Scoring to Prioritization

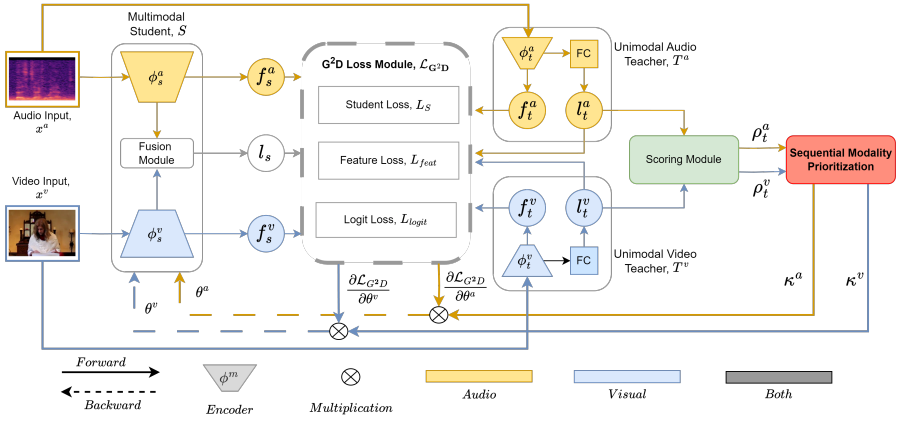


Figure: Now that the **Scoring Module** has generated confidence scores (ρ_t^a , ρ_t^v), we'll see how the **Sequential Modality Prioritization (SMP)** module uses them to modulate the student's gradients.

We propose SMP, a 4-step process to mitigate imbalance, guided by this hypothesis:

Hypothesis 1

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

1. Rank Modalities:

At each training iteration, we rank all modalities from *least confident* ($\pi_t[1]$) to *most confident* ($\pi_t[k]$) based on their unimodal teacher scores (ρ_t^m).

2. The Prioritization Schedule:

- Next, we create a schedule that dedicates a specific number of epochs (τ_j) to training a set of prioritized modalities, $\mathcal{M}_q$.
- This schedule starts by training only the weakest modality, then the second weakest, and so on, before finally training all modalities jointly.

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

3. Modulate Gradients: A modulation coefficient, κ_q^m, acts as a gate, "turning on" gradients only for the modalities currently prioritized in the schedule ($\mathcal{M}_q$).

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

4. Update Student Parameters: This coefficient is applied directly in the gradient update step, effectively zeroing out the updates for non-prioritized modalities.

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

Methodology: The G²D Process Recap

Tying It All Together

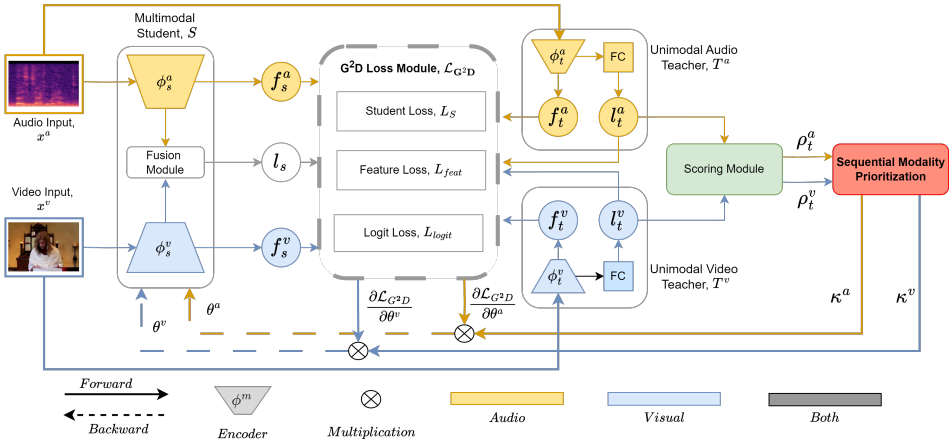


Figure: The complete G²D framework. The student learns via the G^2D **Loss**, guided by stable **teacher confidence scores** that drive the **SMP** mechanism to ensure balanced, interference-free training.

We evaluate G^2D on six diverse, real-world datasets:

Classification Datasets (5 total)

- **CREMA-D** *[Cao et al., 2014]*: An *Audio-Visual* dataset for emotion recognition.
- **AV-MNIST** *[Vielzeuf et al., 2019]*: A synthetic *Audio-Visual* dataset for digit classification.
- **VGGSound** *[Chen et al., 2020]*: A large-scale *Audio-Visual* dataset for event classification.
- **UR-Funny** *[Hasan et al., 2019]*: An *Audio-Visual-Text* dataset for humor detection.
- **IEMOCAP** *[Busso et al., 2008]*: An *Audio-Visual-Text* dataset for emotion recognition.

Regression Dataset (1 total)

- **MIS-ME** *[Rakib et al., 2024]*: An *Image-Tabular* dataset for soil moisture estimation, representing a novel task for evaluating modality imbalance.

Baselines & Backbones

- We compare G^2D against **ten state-of-the-art** baseline methods.
- For a fair comparison, all models use identical backbone architectures:
  - **ResNet-18** *[He et al., 2016]*: For Audio-Visual datasets (CREMA-D, AV-MNIST, VGGSound).
  - **Transformer** *[Vaswani et al., 2017]*: For Audio-Visual-Text datasets (UR-Funny, IEMOCAP).
  - **MobileNetV2** *[Sandler et al., 2018]* **& FCN:** For the Image-Tabular dataset (MIS-ME).

Training & Hyperparameters

- **Fusion Strategy: Late Fusion** *[Gunes and Piccardi, 2005]* is used across all models to ensure a fair comparison.
- **Optimizer:** SGD with a batch size of 16.
- **Hardware:** All models were trained on 3 NVIDIA A10 GPU.
- G^2D **Loss Weights:** For all experiments, the loss weights were set to $\alpha = 1.0$ and $\beta = 1.0$.

Results

Multimodal Performance on Audio-Visual Datasets (Accuracy %)

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Dataset	T^a	T^v	Joint	MSES	MSLR	AGM	PMR	OGM-GE	MLA	MMPareto	ReconBoost	DLMG	UMT	**G²D (Ours)**
CREMA-D	61.7	76.5	67.5	61.0	64.4	78.5	59.1	72.2	79.7	75.1	79.8	<u>83.6</u>	67.6	**85.9**
AV-MNIST	42.7	65.4	69.8	70.7	70.6	72.1	71.8	71.1	65.3	<u>72.6</u>	72.5	72.1	72.3	**73.0**
VGGSound	43.4	32.3	51.0	50.8	51.0	47.1	33.1	51.5	51.7	49.7	51.0	52.7	<u>53.7</u>	**53.8**

Key Findings:

- **Modality imbalance is dataset-dependent**: On CREMA-D the *video* modality (27%) is suppressed, while on AV-MNIST the *audio* modality (16%) is the weaker one.
- G^2D **surpasses all baselines**: Our method consistently achieves the best performance, showing that the SMP strategy ensures more balanced optimization and superior multimodal integration.
- G^2D **outperforms the competing KD-based method**: The results show that our unique loss and dynamic training strategy outperform the UMT baseline ([Du et al., 2023]) across all datasets.

Results

Three-Modality (UR-Funny) & Regression (MIS-ME) Performance

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Three Modalities (UR-Funny, Acc %)

Modality	Joint	OGM-GE	Recon	UMT	G²D
Audio	55.0	50.3	51.7	50.7	**59.2**
Visual	54.9	55.7	55.3	54.9	**55.9**
Text	58.3	55.7	56.3	52.7	**58.2**
Multi	62.6	63.7	61.4	63.4	**65.5**

Key Findings:

- G^2D excels in three-modality settings, enhancing overall performance and individual contributions.
- It avoids **"modality depression,"** where other methods over-suppress the dominant modality (text).

Regression Task (MIS-ME)

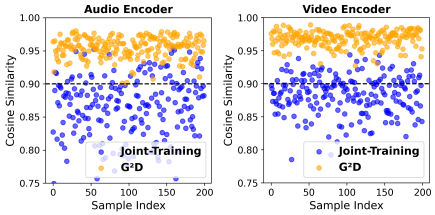
Method	MAPE ↓	R² ↑
Joint-Train	14.62	0.42
MIS-ME *[Rakib et al., 2024]*	7.52	0.80
G²D (Ours)	**7.01**	**0.82**

Key Finding:

- G^2D's versatility is proven by its superior performance on regression, a novel task for imbalance analysis.

Analysis: Feature Alignment

How well do the student's features match the teacher's?



(a) Audio Encoder (b) Video Encoder

Figure: Alignment between unimodal teacher and multimodal student features on CREMA-D, measured by cosine similarity.

Analysis Method:

- We measure the **cosine similarity** between the student's features and the expert unimodal teacher's features on the CREMA-D dataset.

Key Findings:

- The plots show that feature alignment for both audio (left) and video (right) is **consistently higher** with G^2D (orange dots) compared to Joint-Training (blue dots).
- **Conclusion:** This improved feature alignment is a key factor in how G^2D successfully mitigates modality imbalance, ensuring the student learns robustly from each modality.

Analysis: Confidence Ratio

Quantifying the Suppression of Weak Modalities

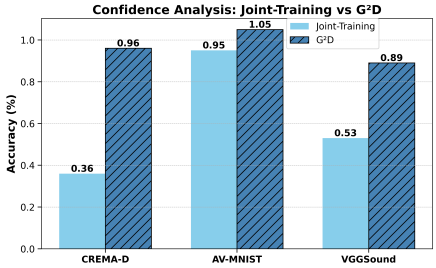


Figure: Confidence ratio of the weaker modality.

Analysis Method:

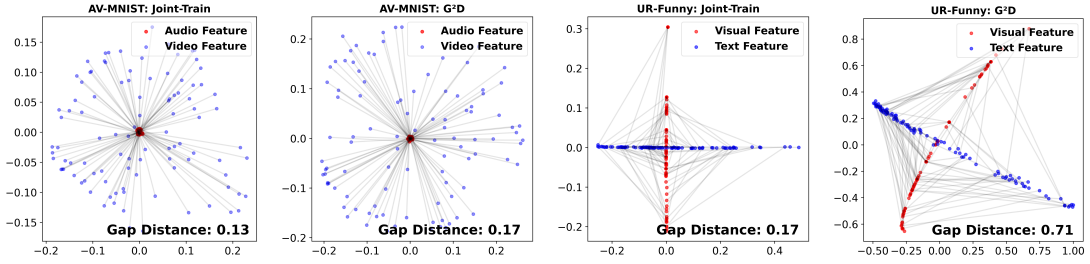
- We define a **Confidence Ratio** to quantify how much a weak modality is suppressed.
- It measures the modality's confidence within the multimodal model, *normalized* by the score of its expert unimodal teacher.
- A **higher ratio** indicates the modality is performing near its full potential; a **lower ratio** indicates suppression.

Key Finding

The bar chart shows that G^2D **consistently yields a much higher confidence ratio** than standard Joint-Training. This demonstrates that our method effectively mitigates modality suppression and ensures a more balanced optimization process.

Analysis: Modality Gap

Visualizing the Separation of Modality Embeddings



(a) AV-MNIST: Joint (b) AV-MNIST: G²D (c) UR-Funny: Joint (d) UR-Funny: G²D

Key Findings

- *[Liang et al., 2022]* shows that a larger *modality gap* (more distinct embeddings) often correlates with better performance.
- G^2D **creates a more pronounced modality gap** than Joint-Training on both datasets, preserving key modality-specific traits that enhance performance.

Ablation Study: Impact of SMP

Is SMP effective on its own?

Table: Effect of adding SMP to different 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
G²D loss	✗	78.63	72.76	63.78
	✓	**85.89**	**73.03**	**65.49**

Key Findings:

- **SMP has a universal benefit:** Integrating our SMP strategy significantly boosts the performance of not only our method, but also standard *Joint-Training* and the competing *UMT* baseline.
- **Synergy with G^2D Loss:** The combination of our proposed G^2D loss with SMP achieves the best overall performance, confirming the effectiveness and synergy of our framework's components.

Ablation Study: G²D with Various Fusion Modules

Does the choice of fusion strategy matter?

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Table: G²D performance with different fusion strategies (Accuracy %).

Fusion Strategy	CREMA-D	AV-MNIST	VGGSound	UR-Funny
Sum	81.59	72.70	50.67	63.08
Concat	83.60	<u>72.98</u>	53.40	64.49
FiLM *[Perez et al., 2018]*	84.27	72.73	48.11	63.48
BiGated *[Kiela et al., 2018]*	81.32	72.89	46.66	63.38
Cross-Attention *[Chen et al., 2021]*	<u>85.35</u>	72.96	<u>53.58</u>	<u>65.09</u>
Late Fusion ***[Gunes and Piccardi, 2005]***	**85.89**	**73.03**	**53.82**	**65.49**

Key Findings

- **Late Fusion achieves the best results**, highlighting the effectiveness of preserving the independent representations from each modality.
- **Cross-Attention is a very close second**, demonstrating its strength in modeling and enhancing cross-modal interactions.

Ablation Study: Modality Suppression in G²D

Partial vs. Complete Suppression

Table: Comparing partial gradient reduction vs. complete gradient shutdown (Accuracy %).

Suppression Type	CREMA-D	AV-MNIST	VGGSound	UR-Funny
Partial (OGM-GE *[Peng et al., 2022]*)	81.99	72.83	51.16	63.68
Complete (SMP)	**85.89**	**73.03**	**53.82**	**65.49**

Key Findings

- **Partial Suppression**, which follows OGM-GE, reduces the gradients of dominant modalities but still allows them to train, meaning modality competition can persist.
- **Complete Suppression** (SMP) completely zeroes out the gradients for non-prioritized modalities.
- **Complete suppression consistently outperforms partial suppression** by allowing the weaker modality to train in isolation, reducing interference and enhancing learning.

Ablation Study: Effect of Prioritization Epochs (τ_j)

How much dedicated training time do weak modalities need?

Table: Two Modalities, Acc (%)

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

Table: Three Modalities, Acc (%)

(τ_1, τ_2, τ_3)	(0,0,150)	(50,50,150)	(75,75,150)
IEMOCAP	75.30	<u>76.99</u>	**77.19**

Key Findings

- The schedule $(\tau_1, \tau_2, \dots)$ defines the number of epochs for prioritizing the weakest modality, then the second weakest, and so on, before a final joint training phase.
- Results show that **increasing dedicated training epochs for weaker modalities improves performance** in both two and three-modality datasets.
- This finding **validates Hypothesis 1:** *interference-free* training time for *weaker* modalities is crucial for mitigating modality imbalance.

Conclusion & Future Work

Summary of Contributions and Broader Impact

In this work, we addressed the challenge of **modality imbalance** in multimodal learning.

Summary of Contributions:

- **Introduced a novel framework,** G^2D that combines *Gradient-Guided Distillation* with *Sequential Modality Prioritization (SMP)* to ensure all modalities contribute effectively during training.
- **Outperformed 10 SOTA baselines** across *six diverse datasets*, including both classification and regression tasks.
- **Successfully mitigated modality imbalance** by dynamically prioritizing and boosting weaker modalities (validated by confidence ratio and feature alignment analysis).

Future Impact

Holds great potential to advance balanced learning in complex multimodal scenarios, paving the way for more inclusive and robust AI systems.

References I

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

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

Chen, C.-F. R., Fan, Q., and Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356.

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

Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., and Zhao, H. (2023). On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8632–8656. PMLR.

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

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

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

Hasan, M. K., Rahman, W., Bagher Zadeh, A., Zhong, J., Tanveer, M. I., Morency, L.-P., and Hoque, M. E. (2019). 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. Association for Computational Linguistics.

References II

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kiela, D., Grave, E., Joulin, A., and Mikolov, T. (2018). 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'18/IAAI'18/EAAI'18. AAAI Press.

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

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. (2022). 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.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. (2018). 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'18/IAAI'18/EAAI'18. AAAI Press.

Rakib, M., Mohammed, A. A., Diggins, D. C., Sharma, S., Sadler, J. M., Ochsner, T., and Bagavathi, A. (2024). Mis-me: A multi-modal framework for soil moisture estimation.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE CVPR*, pages 4510–4520.

References III

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. (2019). 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. Springer-Verlag.

Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., and Carneiro, G. (2023). Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023*, pages 216–226.

Yao, Y. and Mihalcea, R. (2022). Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834.

Zhang, X., Yoon, J., Bansal, M., and Yao, H. (2024). Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27456–27466.

Thank You!

Questions?

Contact: *mohammed.rakib@okstate.edu*