Revisit Modality Imbalance at the Decision Layer

Xiaoyu Ma [1] [2] **Hao Chen** [1] [2]

Abstract

Multimodal learning integrates information from different modalities to enhance model performance, yet it often suffers from modality imbalance, where dominant modalities overshadow weaker ones during joint optimization. This paper reveals that such an imbalance not only occurs during representation learning but also manifests significantly at the decision layer. Experiments on audio-visual datasets (CREMAD and Kinetic-Sounds) show that even after extensive pretraining and balanced optimization, models still exhibit systematic bias toward certain modalities, such as audio. Further analysis demonstrates that this bias originates from intrinsic disparities in feature-space and decision-weight distributions rather than from optimization dynamics alone. We argue that aggregating uncalibrated modality outputs at the fusion stage leads to biased decision-layer weighting, hindering weaker modalities from contributing effectively. To address this, we propose that future multimodal systems should focus more on incorporate adaptive weight allocation mechanisms at the decision layer, enabling relative balanced according to the capabilities of each modality.

1. Balanced Multimodal Learning

In an ideal multimodal model, each modality should be fully optimized during training, and the model should be capable of intelligently integrating the contributions of each modality during decision-making. However, existing studies have revealed the phenomenon of modality laziness (Wang et al., 2020), where certain modalities fail to be sufficiently learned when the model is optimized with a unified objective. This issue is often attributed to *Modality Imbalance* (Peng et al., 2022) during training. Specifically, due to the Greedy

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China [2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China. Correspondence to: Hao Chen <haochen303@seu.edu.cn>.
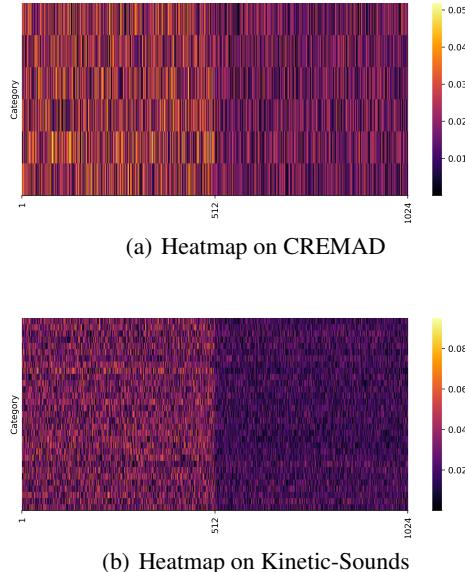
(a) Heatmap on CREMAD

(b) Heatmap on Kinetic-Sounds

Figure 1. **Heatmap of decision-layer weights for a joint-trained multimodal model.** The model employs concatenation for feature fusion, with the first 512 dimensions corresponding to audio features and the last 512 dimensions corresponding to video features.

nature of deep models in joint optimization (Wu et al., 2022), the model tends to prioritize modalities that are easier to learn (the strong modalities), which suppresses the learning of more challenging ones (the weak modalities).

To address this problem, research in the field of *Balanced Multimodal Learning(BML)* (Xu et al., 2025) has explored various strategies to ensure that all modality encoders are optimized in a more balanced manner. Some works mitigate modality laziness by designing modality-specific optimization objectives for the weak modality, such as adjusting task supervision (Wang et al., 2020; Du et al., 2021; Fan et al., 2023; Wei & Hu, 2024), introducing prototype learning (Fan et al., 2023), or employing knowledge distillation (Du et al., 2021). Another line of research aligns optimization dynamics across modalities by adaptively adjusting learning rates (Sun et al., 2021; Peng et al., 2022; Li et al., 2023) or gradient updates (Kontras et al., 2024; Guo et al., 2024) based on unimodal performance. Other methods attempt to decouple the multimodal learning process into alternating

unimodal learning phases (Wu et al., 2022; Zhang et al., 2024; Ma et al., 2025). These methods enhance the learning of individual modality encoders and promote balanced optimization; however, they overlook the fact that **modality imbalance exists not only in the learning capacity of encoders but also in decision-making**, where the model exhibits significant bias during modality fusion.

2. Observation of Modality Imbalance at Decision Layer

In multimodal models, the final decision should be distributed in the fusion layer according to the importance of each modality. To verify this notion, we conducted experiments on two commonly used audio-visual datasets. For the jointly trained multimodal models, we measured the decision-layer weights (L1 norms) and visualized them as heatmaps in Figure 1.

The observations reveal a consistent bias across both datasets: the models tend to rely predominantly on features from the audio modality when making final predictions. This phenomenon closely resembles the modality imbalance frequently discussed in the balanced multimodal learning literature, suggesting that it represents a decision-level manifestation of modality imbalance. At the same time, this behavior appears to align with the theoretical expectation that the strong modality with higher predictive performance should receive higher decision weights. But is this truly the case?

3. Analysis of Modality Imbalance at Decision Layer

Some researches (Xu et al., 2023; Zong et al., 2024) also observe a phenomenon similar to that shown in Figure 1 and attributed it to modality imbalance arising during the optimization process. Xu et al. (2023) propose MMCosine, a method designed to facilitate the learning of weak modalities by emphasizing the directional alignment of gradients while being independent of their magnitude. However, can the bias in decision-layer weights truly be explained solely by differences in optimization rates across modalities? After existing BML methods have achieved an optimization balance, do the decision-layer weights also become balanced? Moreover, is a completely balanced decision layer necessarily desirable? To address these questions, we conducted a series of in-depth experiments.

3.1. Possible Causes of Modality Imbalance at the Decision Layer

Existing studies attribute such modality imbalance at the decision layer to differences in modality optimization rates,
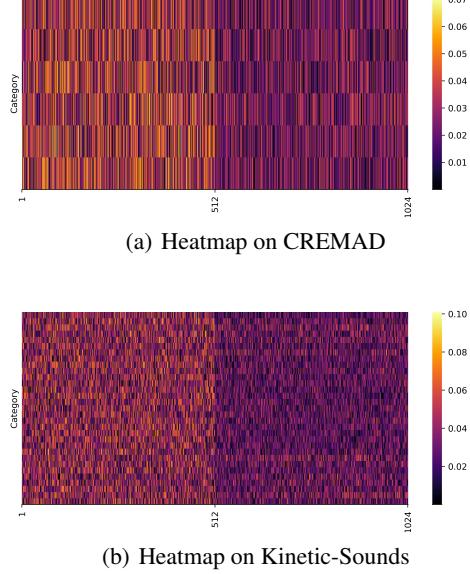

(a) Heatmap on CREMAD

(b) Heatmap on Kinetic-Sounds

Figure 2. **Heatmap of decision-layer weights for a pretrain-based multimodal model.** The model employs concatenation for feature fusion, with the first 512 dimensions corresponding to audio features and the last 512 dimensions corresponding to video features.

where the strong modality dominates the learning process, leading to biased decision-layer weights and logits. To verify the validity of this explanation, we construct a modality-sufficient multimodal model. Specifically, each modality is first fully pre-trained independently, after which the pre-trained weights are transferred into a multimodal framework. Then, only the decision layer is fine-tuned to achieve feature fusion and final prediction. Although this approach is computationally inefficient, it represents an extreme case that ideally satisfies the expectation of balanced modality learning.

However, as shown in Figure 2, visualizing the decision-layer weights reveals that the modality bias remains significant, even after sufficient pre-training. This observation suggests that attributing the imbalance in decision-layer weights solely to differences in optimization rates is not sufficient, and we get our first insight as follows:

> **Insight 1**
>
> *Modality imbalance at the decision layer is not merely caused by differences in modality optimization rates, and BML methods that focus on enhancing encoder capabilities cannot address it.*

Based on the above results, we hypothesize that the observed bias originates from the inherent differences between

(a) Audio - CREMAD
(b) Video - CREMAD

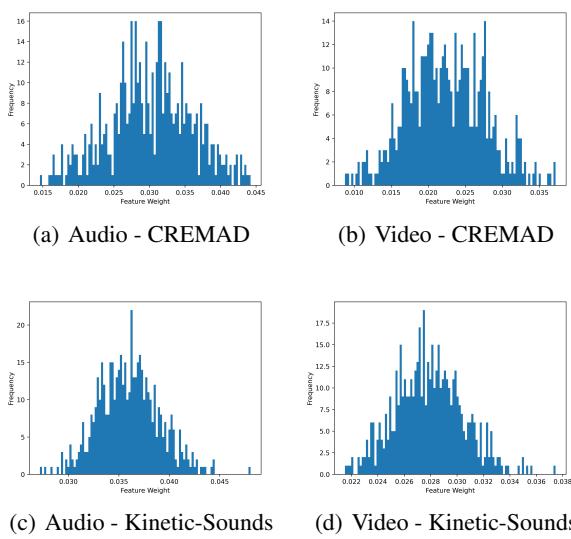(c) Audio - Kinetic-Sounds
(d) Video - Kinetic-Sounds

Figure 3. Distribution of decision-layer weights for different unimodal models trained on audiovisual datasets.

modalities themselves. Specifically, due to the discrepancy in feature-space distributions, each modality tends to learn decision weights following its own intrinsic pattern. To verify this hypothesis, we trained unimodal models separately on both datasets and measured the distribution of their decision-layer weights. As shown in Figure 3, the mean weight magnitude of the audio modality is notably higher than that of the video modality even under unimodal training. This finding supports our assumption that such weight disparities are determined by the intrinsic properties of each modality, rather than being purely an artifact of joint optimization.

Method	CREMAD		Kinetic-Sounds	
	Weight	Logits	Weight	Logits
Audio	3.56	2.14	3.63	2.47
Video	1.81	1.48	2.73	2.02
Multi - Audio	2.13	1.89	3.01	2.83
Multi - Video	1.55	0.58	1.87	1.43

Table 1. Mean values of Weights ($\times 10^{-2}$) and Logits across modalities on the CREMAD and Kinetic-Sounds datasets.

To illustrate this discrepancy more clearly, we summarize in Table 1 the mean values of both the decision-layer weights and the output logits. It can be observed that the imbalance in weights further affects the range of the logits. While such range differences have little impact on unimodal prediction due to the subsequent softmax normalization, they can cause bias in the decision layer of multimodal models when concatenation is used for feature fusion. Without proper cor-
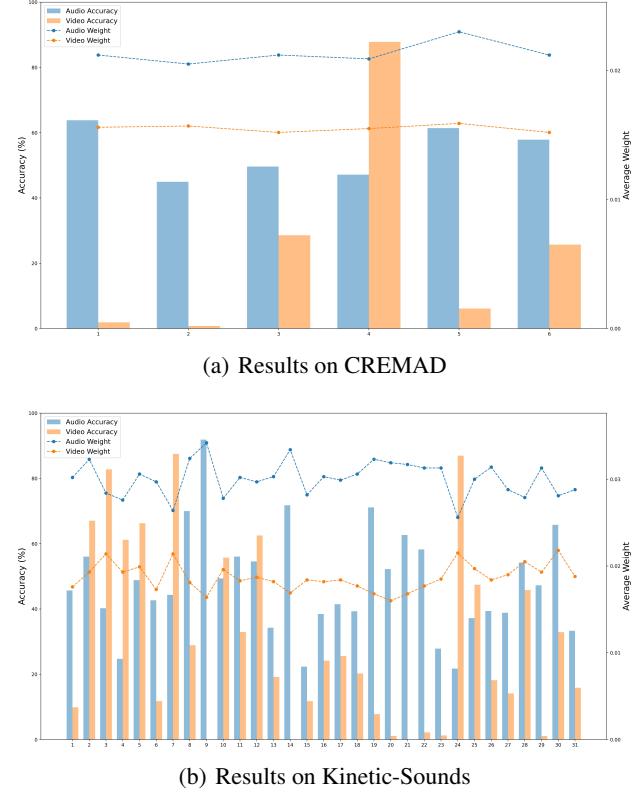

(a) Results on CREMAD

(b) Results on Kinetic-Sounds

Figure 4. **Per-Category Accuracy and Modality Weight at the Decision Layer.** The bar chart represents the unimodal prediction accuracy of each modality across categories, while the line chart illustrates the average L1-norm of decision-layer weights for each modality in different categories.

rection, directly aggregating these imbalanced logits leads to biased decision weights, preventing certain modalities from fully contributing to the final prediction. The above experiments support our second insight:

> **Insight 2**
>
> *The bias in decision-layer weights originates from the inherent differences in modality data distributions, and the Modality Imbalance at the decision-layer is an incorrect retention of this phenomenon.*

3.2. Should We Pursue Decision-Layer Balance?

Existing methods enhance encoder capability by promoting balance at the decision layer and logits level. However, such approaches rely on a fundamental assumption that the contribution of each modality at the decision layer aligns with its performance. Yet, does this assumption truly hold? At the aggregate level, it appears consistent: the audio modality, which dominates the learning process and thus acquires

stronger encoding capability, tends to receive higher decision weights. However, in a classification task, the decision process for each category is relatively independent. Therefore, we further evaluated the modality performance and the average decision-layer weights for each category, as shown in Figure 4. The results reveal a significant variation in discriminative ability across categories, yet the decision weights consistently exhibit a bias toward the audio modality. This observation contradicts the common expectation that a multimodal model should assign larger decision weights to the strong modalities according to their predictive capability. Hence, we propose our third insight:

> **Insight 3**
>
> *Modality Imbalance at the decision layer makes the model cannot automatically allocate decision-layer weights to match the capability of each modality.*

To address this imbalance, previous studies have primarily focused on promoting alignment at the modality level, thereby emphasizing equal contribution across modalities. However, as shown in Figure 4, different modalities exhibit distinct decision capabilities across categories, and the decision processes are relatively independent and adjustable. Therefore, we argue that such adjustment should occur at the category level and should reflect a capability-aware relative balance, which constitutes our fourth insight:

> **Insight 4**
>
> *At the decision layer, relative balance should be promoted at the task level (e.g., per category) according to the capabilities of each modality.*

4. Conclusion

In this reports, we argue that modality imbalance at the decision layer is a highly significant yet long-overlooked problem. It arises not only from differences in optimization rates during training but also from inherent disparities in the feature and decision-weight distributions of different modalities. This phenomenon prevents multimodal models from fully leveraging the strengths of each modality, thereby limiting overall performance. Addressing this issue requires more than merely aligning decision-layer weights; it necessitates identifying the modalities that contribute most effectively and optimizing their decision weights accordingly, enabling the model to adaptively adjust weight allocation based on modality capabilities.

References

Du, C., Li, T., Liu, Y., Wen, Z., Hua, T., Wang, Y., and Zhao, H. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021.

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

Guo, Z., Jin, T., Chen, J., and Zhao, Z. Classifier-guided gradient modulation for enhanced multimodal learning. *arXiv preprint arXiv:2411.01409*, 2024.

Kontras, K., Chatzichristos, C., Blaschko, M., and De Vos, M. Improving multimodal learning with multi-loss gradient modulation. *arXiv preprint arXiv:2405.07930*, 2024.

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

Ma, X., Chen, H., and Deng, Y. Improving multimodal learning balance and sufficiency through data remixing. *arXiv preprint arXiv:2506.11550*, 2025.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

Sun, Y., Mai, S., and Hu, H. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021.

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

Wei, Y. and Hu, D. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *International Conference on Machine Learning*, pp. 52559–52572. PMLR, 2024.

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

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

Xu, S., Cui, M., Huang, C., Wang, H., and Hu, D. Balancebenchmark: A survey for multimodal imbalance learning. *arXiv preprint arXiv:2502.10816*, 2025.

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

Zong, D., Ding, C., Li, B., Li, J., and Zheng, K. Balancing multimodal learning via online logit modulation. In Larson, K. (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5753–5761. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/636. URL https://doi.org/10.24963/ijcai.2024/636. Main Track.