See-Saw Modality Balance: See Gradient, and Sew Impaired Vision-Language Balance to Mitigate Dominant Modality Bias

JuneHyoung Kwon[1*], **MiHyeon Kim**[1*†], **Eunju Lee**[2], **Juhwan Choi**[1], **YoungBin Kim**[1,2]

[1] Department of Artificial Intelligence, Chung-Ang University
[2] Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
{dirchdmltnv, mh10967, dmswn5829, gold5230, ybkim85}@cau.ac.kr

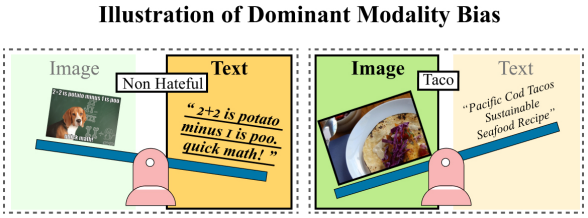
Abstract

Vision-language (VL) models have demonstrated strong performance across various tasks. However, these models often rely on a specific modality for predictions, leading to "dominant modality bias." This bias significantly hurts performance, especially when one modality is impaired. In this study, we analyze model behavior under dominant modality bias and theoretically show that unaligned gradients or differences in gradient *magnitudes* prevent balanced convergence of the loss. Based on these findings, we propose a novel framework, **BALGRAD** to mitigate dominant modality bias. Our approach includes inter-modality gradient reweighting, adjusting the gradient of KL divergence based on each modality's contribution, and inter-task gradient projection to align task *directions* in a non-conflicting manner. Experiments on UPMC Food-101, Hateful Memes, and MM-IMDb datasets confirm that BALGRAD effectively alleviates over-reliance on specific modalities when making predictions.

1 Introduction

Vision-language (VL) models combine image and text modalities, resulting in powerful multi-modal representations. Owing to this integration of two modalities, these models can achieve higher performance in vision-language tasks. Recently, leveraging extensive datasets, VL models have demonstrated remarkable performance across various tasks such as image captioning (Hu et al., 2022), visual question answering (Khademi et al., 2023), and cross-modal retrieval (Liu et al., 2023), showcasing their capability to harness the complementary strengths of visual and textual data.

However, these models often rely on a single modality rather than treating and utilizing them equally, leading to the dominance of a certain modality on the overall performance. A conceptual overview of this effect can be seen in Figure 1. This phenomenon, where a specific modality disproportionately influences the model's outcomes, is referred to as "dominant modality bias" (Woo et al., 2023). For instance, VL models tend to be biased towards the text modality when recognizing hate expressions (Kiela et al., 2020; Aggarwal et al., 2024), thereby limiting the VL model's ability to effectively integrate and interpret images.



(a) Hate Recongition (b) Food Classification

Figure 1: Conceptual visualization of dominant modality bias. The key modality differs by task: **(a)** For the hate recognition task, text descriptions of memes lead, while **(b)** for the food classification task, food images play a crucial role in prediction.

This bias behaves particularly detrimentally when one modality is impaired, such as when data is noisy and it is difficult to gather paired data (Garg et al., 2022; Woo et al., 2023; Yang et al., 2024). This issue is common in real-world scenarios due to privacy-related data sharing restrictions or stringent data storage policies (Voigt and Von dem Bussche, 2017) and can severely degrade the model's performance. Additionally, the failure to sufficiently explore the weak modality limits the overall performance of the VL model (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022), highlighting the need for robust solutions to mitigate dominant modality bias.

To address this issue and balance the information between modalities, numerous studies have

[*] Equal contribution.
[†] Currently at: KT CORPORATION, mihyeon.gim@kt.com.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4364–4378

been conducted. Several studies have focused on modulating the gradients of each encoder based on the confidence of individual modalities (Peng et al., 2022; Li et al., 2023). Other approaches have involved training multimodal models using the best-performing learning rates from unimodal models (Yao and Mihalcea, 2022). However, these methods often induce negative transfer (Wang et al., 2019; Yu et al., 2020), which occurs when the model's performance decreases with the addition of modality data compared to solely using unimodal data.

We first analyze the behavior of models after the dominant modality bias has taken root. Our analysis reveals that certain modalities are more crucial to target performance and observes that the dominant and weak modalities converge at different rates during training. Additionally, we theoretically demonstrate that the balanced convergence of the loss is influenced by both the *magnitude* and *direction* of the gradient. Based on these findings, we propose **BALGRAD (Balancing Gradients)** to mitigate dominant modality bias. Firstly, we adopt a mutual KL divergence between the predictions of each modality to ensure balanced updates. However, a naive approach that equally aligns the distributions of two modalities can hinder the representation learning of each modality. To address this, we introduce **inter-modality gradient reweighting**, which adjusts the *magnitude* of the gradient of the KL divergence term based on the learning status of each modality. Additionally, we propose **inter-task gradient projection**, which updates the gradient of the target task to establish a balance between both modalities. We project the target task's gradient in a *direction* orthogonal to the KL divergence gradient if a conflict between the gradients occurs, encouraging stabilized training between the two modalities.

We evaluate the effectiveness of BALGRAD on models using three vision-language datasets: UPMC Food-101 (Wang et al., 2015), Hateful Memes (Kiela et al., 2020), and MM-IMDb (Arevalo et al., 2017). To simulate the influence of individual modalities, we conduct experiments under conditions where specific modalities are missing or impaired by noise. The experimental results demonstrate that the proposed method reduces the gap between the modalities while avoiding negative transfer. The contributions of our proposed method are as follows:

- We analyze the dominant modality bias and theoretically demonstrate that the balanced convergence of loss is influenced by both the *magnitude* and *direction* of the gradient.
- We propose BALGRAD, which reweights the gradients between modalities to ensure stable convergence and projects the target task's gradient to avoid conflicts that hinder balanced learning.
- Experimental results across UPMC Food-101, Hateful Memes, and MM-IMDb under different impaired conditions confirm the effectiveness of our proposed method in mitigating dominant modality bias.

2 Related Work

In multimodal models, such as VL models, a bias towards a preferred or easier-to-learn modality often leads to the under-exploration of others (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022). Studies have analyzed this, noting that multimodal models are prone to overfitting and show discrepancies in generalization across modalities (Wang et al., 2020). Differences in convergence speeds also contribute to this bias (Yao and Mihalcea, 2022; Wu et al., 2022). An early study in this field finds that certain modalities, correlating with their network's random initialization, dominate the learning process (Huang et al., 2022), while other researchers attribute the preference to unimodal representation margins and insufficient integration of modalities (Yang et al., 2024). Another line of study highlights that spurious correlations with instance labels cause imbalances in modality utilization (Guo et al., 2023). In this paper, we identify that the dominant modality bias in VL models arises from the influence of gradient magnitude and direction on the model's loss function, hindering balanced learning across modalities.

In response to the challenge of balancing modalities in multimodal learning, various strategies have been proposed. MSLR suggests using different optimal learning rates for each modality during multimodal learning to enhance performance (Yao and Mihalcea, 2022). Another approach involves using a conditional utilization rate to re-scale modality features, ensuring balanced contributions from each modality (Wu et al., 2022). Gradient blending optimizes the mixing of modalities based on the model's overfitting behavior (Wang et al.,

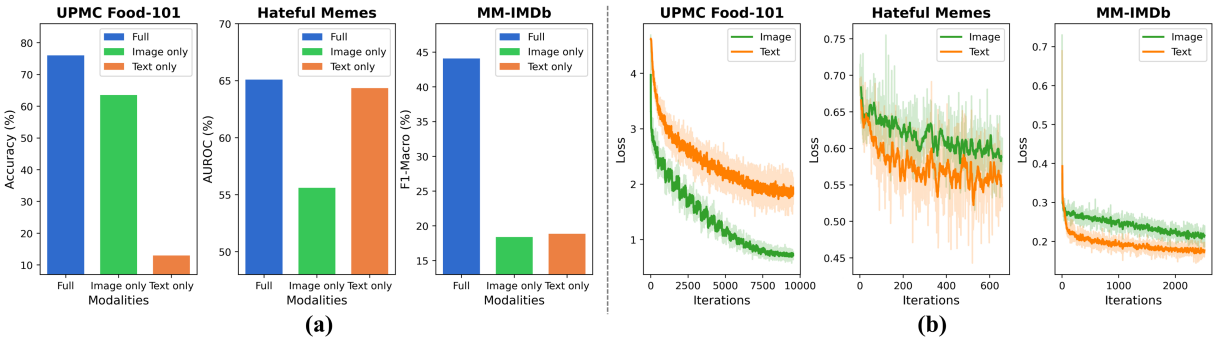


Figure 2: Experimental results on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets in the presence of dominant modality bias. **(a)** Performance visualization under different missing conditions (full, image only (missing text), text only (missing image)) for each dataset. **(b)** Illustration of learning curves for each modality across datasets.

2020). OGM-GE adaptively controls the optimization process using modality-specific confidence scores (Peng et al., 2022). AGM employs Shapley values to modulate gradients through mono-modal responses, aiming to balance the learning process across modalities (Li et al., 2023). However, these methods often lack consideration of negative transfer and may introduce adverse effects. In this paper, we propose BALGRAD, which reweights gradients considering the learning status of each modality and projects the gradients to mitigate dominant modality bias without disrupting the balance between modalities.

3 Method

In this section, we analyze the dominant modality bias and propose BALGRAD to mitigate such bias. In Section 3.1, we observe the behavior of VL models and theoretically demonstrate the factors influencing balanced loss convergence. In Section 3.2, based on these findings, we introduce BALGRAD, which reweights and projects gradients to ensure balanced learning across modalities.

3.1 Analysis of Dominant Modality Bias

We introduce a controlled experiment to analyze the behavior of VL models biased by dominant modality. We denote the training dataset as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = (x_i^v, x_i^l)$ is a pair of data from the image and text modalities, respectively, and y_i represents the label. We extract features from the image and text encoders, passing them through their respective embedding layers, $h_v(\cdot)$ and $h_l(\cdot)$. These embeddings are then fused via concatenation and passed through a classifier, $f_{\mathcal{T}}(\cdot)$, to yield the predicted probability $p_{\mathcal{T}}$. Details on the architecture and training scheme are provided in the Appendix B.

Analysis on Performance Gap. To analyze the impact of individual modalities on the performance of VL models, we mute one modality by inputting empty values at the data level, rendering it non-informative. This method is applied while testing on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets. The experimental results in Figure 2 (a) show a significant performance drop when a specific modality is missing. In the case of UPMC Food-101, the image modality significantly influences the overall performance, while in Hateful Memes, the text modality plays a more crucial role. Conversely, the performance drop is relatively minor when the weak modality (*text* for *UPMC Food-101* and *image* for *Hateful Memes*) is missing. In contrast, for MM-IMDb, the performance drop is similar when either modality is missing, indicating that the model is not biased towards a specific modality.

Analysis on Training Dynamics. To observe the loss dynamics of each modality during the training phase, we add linear classifiers $f_v(\cdot)$ and $f_l(\cdot)$ on top of the image and text embedding layers, respectively. These classifiers output probabilities p_i^v and p_i^l, which are then used to predict the label y_i, and each target objective is represented as $\mathcal{L}_{\mathcal{T}}^v$ and $\mathcal{L}_{\mathcal{T}}^l$, respectively. We find that the loss of the dominant modality decreases rapidly, while the loss of the weak modality decreases relatively slowly, as shown in Figure 2 (b). For MM-IMDb specifically, the loss gap decreases as training iterations increase, demonstrating that the model is not biased toward any single modality. This indicates that, during training, one modality is overly exploited while the other modality is relatively underexplored, which is consistent with previous re-

search (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022). We conjecture that this phenomenon appears inherently task-dependent, with the VL model inclined to update based on the easy-to-learn modality that can quickly reduce the loss (Arpit et al., 2017; Nam et al., 2020).

Theoretical Analysis of Gradient Influence. To theoretically analyze why VL models struggle to balance the utilization of both modalities, we examine the loss reduction in terms of gradient updates. The loss function for a target is defined as $\mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T})$, where θ_v and θ_l are the parameters of image and text embedding layers, respectively, and the $\theta_\mathcal{T}$ represents the parameters of the classifier $f_\mathcal{T}(\cdot)$. The objective is to find the optimal parameters $\Theta = \{\theta_v, \theta_l, \theta_\mathcal{T}\}$ that minimize $\mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T})$. To analyze how each modality contributes to the overall loss reduction, we decompose the target task loss gradient with respect to the model parameters Θ into modality-specific components, denoted by $\mathcal{G}^\tau = \{g_l, g_v, g_\mathcal{T}\}$. These partial gradients capture the influence of linguistic, visual, and task-related parameters, respectively, under standard gradient-descent updates. Additionally, $g_\mathcal{T} = \sum_{i\in\{v,l\}} \nabla_{\theta_i} \hat{y} \nabla_{\hat{y}} p_\mathcal{T} \nabla_{p_\mathcal{T}} \mathcal{L} = \sum_{i\in\{v,l\}} g^i_\mathcal{T}$ denotes the gradient for parameters $\theta_\mathcal{T}$ of the linear classifier $f_\mathcal{T}(\cdot)$, where $g^i_\mathcal{T}$ denotes the gradient of each modality in $f_\mathcal{T}(\cdot)$. We theoretically analyze how the target objective is influenced by the varying magnitudes and directions of gradients for each modality.

Proposition 1. (Gradient Effect on Change of Loss) Let the parameters θ_v, θ_l, and $\theta_\mathcal{T}$ of a multimodal model be updated with gradients g_v, g_l, and $g_\mathcal{T}$ using a sufficiently small step size $\lambda > 0$, resulting in updated parameters $\hat{\theta}_v, \hat{\theta}_l$, and $\hat{\theta}_\mathcal{T}$. Then the change in the loss function satisfies

$$\begin{aligned}\triangle\mathcal{L} \;=\; & -2\,\lambda\,(g^v_\mathcal{T} \cdot g^l_\mathcal{T}) \\ & -\,\lambda \sum_{i\in\{v,l,\mathcal{T}\}} \Big(g_i \cdot g_i \;+\; g^i_\mathcal{T} \cdot g^i_\mathcal{T} \Big) \;+\; O(\lambda^2),\end{aligned} \tag{1}$$

where the cross term $-2\,\lambda\,(g^v_\mathcal{T}) \cdot (g^l_\mathcal{T})$ captures the interaction between the visual and language gradients and the magnitudes and directions of each gradient $g^v_\mathcal{T}$ and $g^l_\mathcal{T}$ governs how much the overall loss is reduced.

Proof. See Appendix A.1 □

If the gradients for the two modalities $g^v_\mathcal{T}$ and $g^l_\mathcal{T}$ do not align well, meaning they have conflicting directions or have significantly different magnitudes, the loss reduction will not be balanced. Gradients with larger magnitudes substantially impact loss reduction, while gradients with directions that align more closely between modalities facilitate more effective joint learning. Consequently, the loss is likely to decrease more under the influence of the dominant modality, leading to an uneven contribution from each modality.

3.2 BALGRAD

Based on the findings above, we propose BALGRAD to mitigate the dominant modality bias, which consists of two components: inter-modality gradient reweighting and inter-task gradient projection. Inter-modality gradient reweighting addresses the imbalance caused by different gradient *magnitudes*, ensuring more equal contributions from each modality. Inter-task gradient projection aligns the gradient *directions* of the modalities, facilitating more effective joint learning and preventing the dominant modality from disproportionately influencing loss reduction. The overall process of BALGRAD can be seen in Figure 3.

3.2.1 Inter-modality Gradient Reweighting

Standard VL models lack the consideration to ensure that both modalities are updated equally, leading to the stronger modality dominating the training phase, as we observed in the previous section. Therefore, inspired by knowledge distillation (Hinton et al., 2014; Zhang et al., 2018; Phuong and Lampert, 2019), we aim to balance the gradients received from each modality by aligning the distributions of their predictions. To achieve this, we compute the mutual Kullback-Leibler (KL) divergence between the predictions p^v_i and p^l_i of the two modalities. This involves aligning the predictions of the image modality with those of the text modality and vice versa. The KL divergence from p^l_i to p^v_i is as follows:

$$\mathcal{L}^l_{kl} = -\sum_i p^l_i log \frac{p^v_i}{p^l_i} \tag{2}$$

We also compute $\mathcal{L}^v_{kl}$ in the same manner. We represent the gradients of $\mathcal{L}^l_{kl}$ and $\mathcal{L}^v_{kl}$ as $g^l_{kl} = \nabla\mathcal{L}^l_{kl}$ and $g^v_{kl} = \nabla\mathcal{L}^v_{kl}$, respectively. In this way, each modality's embedding layer learns to correctly predict the label and match the probability estimate of other modalities, thereby alleviating the severe imbalance. However, symmetrically aligning the distributions between the two modalities

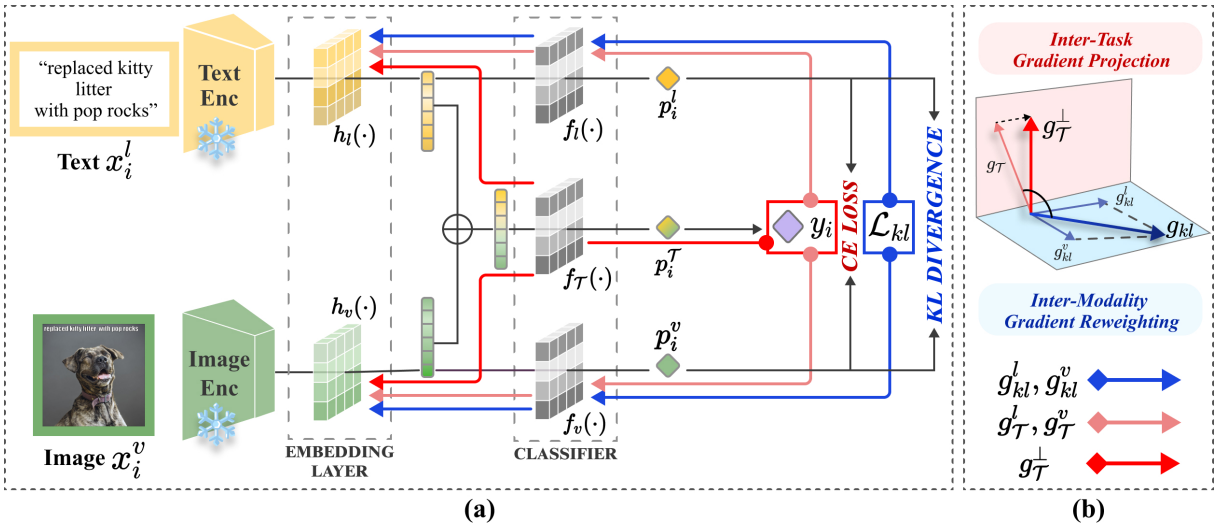


Figure 3: **(a)** The overall training framework of our proposed BALGRAD. The final classifier $f_\mathcal{T}(\cdot)$ is updated with the gradient $g_\mathcal{T}^\perp$ for cross entropy (CE) loss. The image and text embedding layers $h_v(\cdot), h_l(\cdot)$ are also updated with $g_\mathcal{T}^\perp$ along with the gradients of the CE loss for each modality $g_\mathcal{T}^v, g_\mathcal{T}^l$, and the gradients of the KL divergence between the two modalities' predictions g_{kl}^v, g_{kl}^l. **(b)** Inter-modality gradient reweighting adjusts the magnitudes of g_{kl}^v and g_{kl}^l to obtain g_{kl}. If a conflict occurs, we project $g_\mathcal{T}^\perp$ on the orthogonal direction of g_{kl} by inter-task gradient projection.

overlooks the differences in their convergence status, as observed in Section 3.1. This can cause the layers of the faster-converging modality to be hindered in their representation learning, leading to performance degradation. Therefore, we propose an **inter-modality gradient reweighting** method that adjusts the magnitude to which each modality receives the KL divergence gradient based on its contribution to the learning objective. We reweight the gradient of the KL divergence term for p_i^l to p_i^v and p_i^v to p_i^l using the following terms, respectively:

$$\mathcal{W}^l = \frac{\mathcal{L}_\mathcal{T}^l}{\mathcal{L}_\mathcal{T}^v + \mathcal{L}_\mathcal{T}^l}, \ \mathcal{W}^v = \frac{\mathcal{L}_\mathcal{T}^v}{\mathcal{L}_\mathcal{T}^v + \mathcal{L}_\mathcal{T}^l} \quad (3)$$

In this configuration, if the target task loss for a modality is low (i.e., it has converged more), the gradient receives a lower weight. This ensures that the gradient of the weak modality is updated more toward matching the dominant modality's prediction, thereby reducing the training gap. In contrast, the dominant modality receives less influence from the underperforming predictions, allowing it to effectively learn its representation. Additionally, to ensure that each modality is trained for the target task independently, we introduce an additional term that increases the reweighting factor as iteration t progresses. This ensures that the impact of mutual learning grows over time, allowing individual encoders to learn effectively in the initial stages and progressively encouraging balanced learning between modalities. The final reweighted gradient for the KL divergence is as follows:

$$g_{kl} = (\gamma + \frac{\gamma}{1 + e^{-t}})(\mathcal{W}^l g_{kl}^l + \mathcal{W}^v g_{kl}^v) \quad (4)$$

γ is the initial weighting factor, and we set $\gamma = 1/2$.

3.2.2 Inter-task Gradient Projection

Proposition 1 highlights that properly aligning gradients with different directions and magnitudes is crucial for effective joint learning. However, when gradients are not aligned and exhibit negative cosine similarity, known as conflicting gradients, the optimization process becomes suboptimal (Yu et al., 2020; Shi et al., 2023). Such conflicts can arise between the gradients of different tasks, potentially causing the dominant gradient to overwhelm the optimization process at the expense of the other task's performance.

For our case, as confirmed in Section 3.1, the target task, which is $\mathcal{L}_\mathcal{T}$ alone, fails to balance the modalities and fully explore the weak modality. Therefore, we introduce the balance between the predictions of each modality as an additional task. However, as mentioned earlier, naive joint training can cause conflict between the gradient of the target task and KL divergence (i.e., $g_\mathcal{T}$ and g_{kl}).

Proposition 2. (Gradient Conflicts on Loss Reduction with KL Loss) Let $\mathcal{G}^{\tau} = \{ g_v, g_l, g_{\tau}\}$ and $\mathcal{G}^{kl} = \{ g_v^{kl}, g_l^{kl}, 0\}$ be the gradients from a target loss $\mathcal{L}_{\tau}$ and a KL loss $\mathcal{L}_{kl}$, respectively, with parameters $\theta = [\theta_v, \theta_l, \theta_{\tau}]^{\top}$. Assume the parameters are updated by gradient descent with a small step size $\lambda > 0$: $\theta_v' = \theta_v - \lambda \left(g_v + g_v^{kl}\right), \theta_l' = \theta_l - \lambda \left(g_l + g_l^{kl}\right), \theta_{\tau}' = \theta_{\tau} - \lambda\, g_{\tau}$. Then, for the combined loss $\mathcal{L} = \mathcal{L}_{\tau} + \mathcal{L}_{kl}$, the change in the loss is

$$\begin{aligned}\Delta\mathcal{L} &= \mathcal{L}(\theta') - \mathcal{L}(\theta) \\ &= -\lambda\Big(\|\mathcal{G}^{\tau}\|^2 + \|\mathcal{G}^{kl}\|^2 + 2\,(\mathcal{G}^{\tau})^{\top}\mathcal{G}^{kl}\Big) \\ &\quad + O(\lambda^2).\end{aligned} \tag{5}$$

In particular, if $(\mathcal{G}^{\tau})^{\top}\mathcal{G}^{kl} < 0$, the gradients from the target and KL losses *conflict*, reducing the effective loss reduction.

Proof. See Appendix A.2 □

Building upon Proposition 2, we aim to ensure that the gradient of the target task does not disrupt the balance between modalities. Specifically, we propose **inter-task gradient projection**, which projects $g_{\mathcal{T}}$ onto g_{kl} in a non-conflicting manner. First, we consider the relationship between the two gradients to determine if they conflict and compute the cosine similarity between the two gradients. If $g_{\mathcal{T}} \cdot g_{kl} \geq 0$, we assume that $g_{\mathcal{T}}$ is being updated in a direction that aligns with modality balance, and we use the original $g_{\mathcal{T}}$ for updating the model. Conversely, if $g_{\mathcal{T}} \cdot g_{kl} < 0$, indicating a potential disruption to the balance between modalities, we project $g_{\mathcal{T}}$ in a direction orthogonal to g_{kl}. This process can be represented as follows:

$$g_{\mathcal{T}}^{\perp} = \begin{cases} g_{\mathcal{T}} - \left(\frac{g_{\mathcal{T}} \cdot g_{kl}}{\|g_{kl}\|^2}\right) g_{kl}, & \text{if } g_{\mathcal{T}} \cdot g_{kl} < 0 \\ g_{\mathcal{T}}, & \text{otherwise} \end{cases} \tag{6}$$

This projection ensures that $g_{\mathcal{T}}^{\perp}$ is adjusted to maintain the balance between the modalities while preventing conflicts with g_{kl}. In a nutshell, the proposed BALGRAD allows for extensively learning different modalities and tasks, effectively optimizing the target task while maintaining the balance between the modalities.

4 Experiments

4.1 Experimental Setup

We conduct experiments on three vision-language datasets: UPMC Food-101 (Wang et al., 2015), Hateful Memes (Kiela et al., 2020), and MM-IMDb (Arevalo et al., 2017). For image and text encoding, we utilize ViT (Dosovitskiy et al., 2021) and BERT (Devlin et al., 2019), respectively, employing a late concatenation architecture for final predictions. To minimize extensive fine-tuning, we adopt linear probing, freezing all encoder parameters and training only the embedding and classifier layers. Further implementation details are provided in Appendix B.

To assess the robustness of the VL model against dominant modality bias, we introduced two impaired conditions: missing and noisy. For the missing modality, empty strings were used for text and zero pixels for images (Lee et al., 2023). In the noisy condition, 30% salt and pepper noise is added to images (Lim et al., 2023), and 15% of text tokens were randomly deleted (Manolache et al., 2021; Yuan et al., 2023). All experiments were conducted with the model trained on unimpaired full modality data, with impairments applied to the entire data of a specific modality during testing. Further implementation details are provided in Appendix C.

4.2 Experimental Results

We train with full modality data and evaluate the performance of the VL model under conditions where one modality is entirely impaired across three datasets, as shown in Table 1. "Full" refers to the scenario where no modalities are impaired during testing. For the impaired cases (missing and noisy), each modality is impaired according to the specified method. "*Avg.*" denotes the average performance when each modality is impaired individually, while "Δ_{Gap}" represents the performance difference between the image-impaired and text-impaired conditions. A smaller Δ_{Gap} indicates a more balanced model that does not overly rely on a single modality, thereby exhibiting less dominant modality bias.

For the UPMC Food-101 dataset, BALGRAD demonstrates the highest performance across all conditions—full, missing image, and missing text. Notably, it improves the performance on the weak modality, text, by 12.5%p compared to the baseline. Additionally, it achieves the highest average performance and exhibits the smallest gap, effectively mitigating bias despite the dominant influence of the image modality. In the noisy condition, our method shows robustness comparable to AGM (Li et al., 2023) and achieves the highest *Avg.*

BALGRAD exhibits the highest performance

Modality		UPMC Food-101					Hateful Memes					MM-IMDb				
		Baseline	MSLR	OGM-GE	AGM	BALGRAD	Baseline	MSLR	OGM-GE	AGM	BALGRAD	Baseline	MSLR	OGM-GE	AGM	BALGRAD
Full		76.01	78.43	77.42	78.93	**80.32**	65.10	65.58	66.70	64.69	**67.35**	**44.09**	**44.09**	42.22	43.93	43.19
Missing	Image	12.99	20.52	13.86	22.60	**25.49**	64.34	66.04	**66.83***	66.25*	65.86	18.85	19.26	**24.48**	17.57	18.81
	Text	63.52	63.00	61.45	63.13	**65.03**	55.60	55.66	57.20	56.20	**57.58**	**18.40**	14.67	12.31	15.46	17.47
	Avg.↑	38.26	41.76	37.66	42.87	**45.26**	59.97	60.85	**62.02**	61.23	61.72	**18.63**	16.97	18.40	16.52	18.14
	Δ_{Gap}↓	50.53	42.48	47.59	40.53	**39.54**	8.74	10.38	9.63	10.05	**8.28**	**0.45**	4.59	12.17	2.11	1.34
Noisy	Image	41.92	52.92	46.50	**56.57**	55.58	63.64	64.21	63.72	61.85	**65.78**	30.89	33.86	35.31	35.73	**37.76**
	Text	67.28	77.71	75.94	77.43	**78.54**	65.09	63.66	**67.16***	63.68	65.60	38.09	**43.00**	40.33	42.66	41.80
	Avg.↑	54.60	65.32	61.22	67.00	**67.06**	64.37	63.94	65.44	62.77	**65.69**	34.49	38.43	37.82	39.20	**39.78**
	Δ_{Gap}↓	25.36	24.79	29.44	**20.86**	22.96	1.45	0.55	3.44	1.83	**0.18**	7.20	9.14	5.02	6.93	**4.04**

Table 1: The experimental result to validate the effectiveness of BALGRAD on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets. The best result in each test dataset is boldfaced, and the second best is presented with underlining. "*Avg.*" represents the average performance under conditions where one of the modalities is impaired (missing or noisy), while "Δ_{Gap}" indicates the performance difference. The value that is displayed in gray* represents a negative transfer. The unit for "Δ_{Gap}" is %p, and the unit for all other values is %.

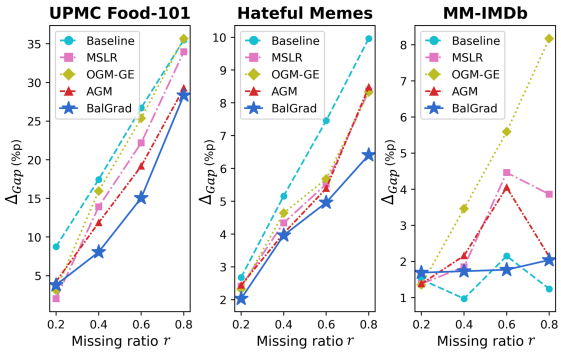


Figure 4: Evaluation on robustness to different missing ratio r of BALGRAD and existing methods on UPMC Food-101, Hateful Memes, and MM-IMDb datasets.

in conditions where the dominant text modality is missing, as well as in the full modality, *Avg.*, and Δ_{Gap} for the Hateful Memes dataset. OGM-GE (Peng et al., 2022) and AGM perform better in the image missing condition than in the full modality condition, indicating a heavy reliance on the text modality, with performance increases of 0.13%p and 1.56%p, respectively. In other words, adding the image modality results in a decrease in performance compared to using text alone, exhibiting negative transfer (Wang et al., 2019). In the noisy condition, BALGRAD demonstrates the highest *Avg.* performance and the smallest Δ_{Gap}, showcasing that BALGRAD sufficiently explores the image modality.

Furthermore, BALGRAD maintains the balance between the modalities even despite the absence of any dominant modality. For the MM-IMDb dataset, our proposed method shows slightly lower performance compared to the baseline but exhibits the second-smallest Δ_{Gap}, indicating balanced results without a dominant modality. Although OGM-GE demonstrates high performance, it exhibits a significant imbalance between modalities, as evidenced by the considerably higher gap, which is 10.83%p more than our method. BALGRAD achieves the highest average performance and the lowest gap in the noisy condition, showcasing that our proposed method effectively explores both modalities without being biased towards one.

Additionally, to investigate the robustness under varying degrees of impairment, we mute a specific modality according to the missing ratio r, and the results are shown in Figure 4. For each dataset, we randomly drop a certain percentage $r\%$ of the data from each modality and measure the resulting performance Δ_{Gap}. We set missing ratios $r \in \{0.2, 0.4, 0.6, 0.8\}$. Experimental results indicate that BALGRAD consistently exhibits a lower gap compared to existing methods across varying missing ratios, demonstrating robustness to impaired modalities. While BALGRAD exhibits a slightly larger gap compared to the baseline, it is noteworthy that BALGRAD significantly reduces the gap for datasets with dominant modality bias. Additionally, it introduces a small gap for datasets where dominant modality bias is not present.

Additional experimental results on various fusion mechanisms, backbone models, and datasets are provided in Appendix C. The results demonstrate that BALGRAD consistently delivers robust performance across different biases, modality types, datasets, and perturbed conditions, underscoring its effectiveness in synergistically integrating modalities to prevent negative transfer and ensure reliable, real-world multimodal learning.

Modality		UPMC Food-101				Hateful Memes				MM-IMDb			
		Baseline	w/ Gradient reweighting	w/ Gradient projection	BALGRAD	Baseline	w/ Gradient reweighting	w/ Gradient projection	BALGRAD	Baseline	w/ Gradient reweighting	w/ Gradient projection	BALGRAD
Full		76.01	78.17	76.20	**80.32**	65.10	65.80	66.30	**67.35**	44.09	**44.30**	42.30	43.19
Missing	Image	12.99	22.30	19.82	**25.49**	64.34	66.37*	65.40	65.86	18.85	**21.48**	18.47	18.81
	Text	63.52	64.10	63.76	**65.03**	55.60	57.03	56.20	**57.48**	18.40	17.20	**18.80**	17.47
	Avg.↑	38.26	43.20	41.79	**45.26**	59.97	**61.70**	60.80	61.67	18.63	**19.34**	18.64	18.14
	Gap	50.53	41.80	43.94	**39.54**	8.74	9.34	9.20	**8.38**	0.45	4.28	**0.33**	1.34

Table 2: Ablation study results compares performance with and without inter-modality gradient reweighting and inter-task gradient projection to evaluate their impact on modality balance and transfer effects on UPMC Food-101, Hateful Memes, and MM-IMDb datasets. The best results are highlighted in bold, the second-best in italics, and values shown in gray* indicate negative transfer. "Δ_{Gap}" is reported in %p, while all other values are in %.

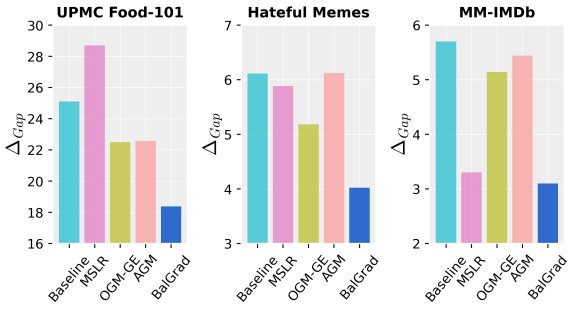


Figure 5: Bar plots comparing the performance of existing methods and BALGRAD using BLIP. Each bar represents $\Delta_{Gap}(\%)$, defined as the performance difference between missing image and missing text conditions.

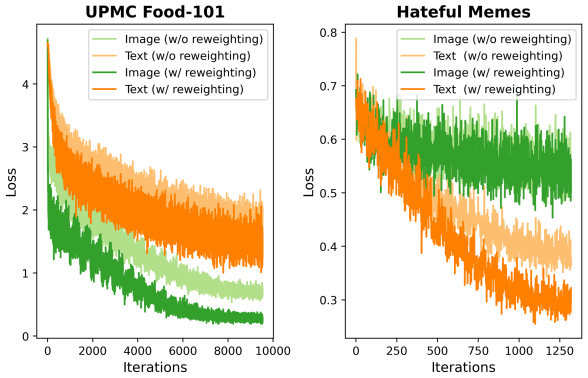


Figure 6: Training iteration loss curves for image and text modalities on the UPMC Food-101 and Hateful Memes datasets, comparing the effects of the existence of inter-modality gradient reweighting.

4.3 Ablation and Analysis

Analysis of Each Component. We conduct ablation experiments to assess the impact of gradient reweighting and projection as shown in Table 2. While gradient reweighting shares a common approach with existing methods (Peng et al., 2022; Li et al., 2023), helps mitigate modality imbalance, it induces negative transfer in the Hateful Memes dataset and leaves the MM-IMDb dataset overly reliant on text. In contrast, incorporating gradient projection eliminates negative transfer and balances modality use. By aligning the gradient of the target loss with the KL loss term, we reduce reliance on any single modality, effectively preventing negative transfer. These points clarify how our approach differs from existing work and address the gaps in empirical validation and mitigation of negative effects.

Evaluation on Text Decoder-based Vision-Language Model. To examine BALGRAD's effectiveness in text decoder-based architectures, we conduct additional experiments using BLIP (Li et al., 2022), which generates textual outputs from visual inputs via a text decoder. This setup differs from encoder-only VL models and aligns with autoregressive language modeling approaches. As shown in Figure 5, BALGRAD achieves the lowest Δ_{Gap} across all datasets, indicating its ability to balance modality contributions in decoder-based VL models. These results highlight BALGRAD's potential for extension to decoder-only LLMs, as it effectively mitigates dominant modality bias across different VL architectures.

Ablation on Inter-modality Gradient Reweighting. To validate the efficacy of inter-modality gradient reweighting, we track the training loss dynamics for each modality on datasets with dominant modality bias (UPMC Food-101 and Hateful Memes), as shown in Figure 6. Without reweighting, weights are fixed at $\mathcal{W}^v = 1/2$ and $\mathcal{W}^l = 1/2$, equally distilling information between modalities. Experimental results show that reweighting leads to faster and more stable convergence of loss for each modality. This supports Proposition 1 in Section 3.1, indicating that gradient reweighting optimizes the exploration of individual modalities while maintaining balance in the VL model.

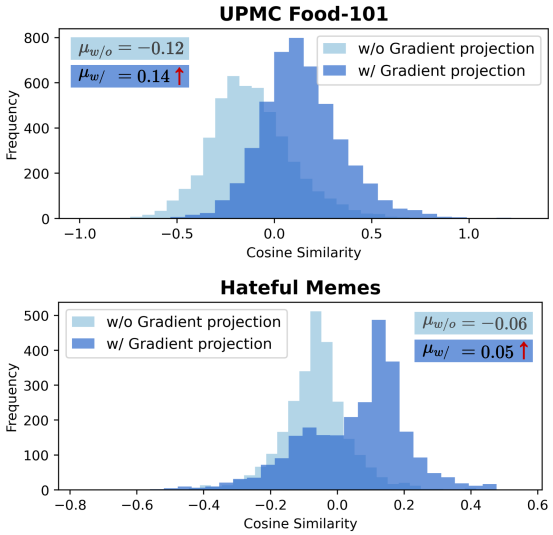


Figure 7: Histogram visualization of the frequency of gradient conflicts between image and text gradients during training iterations on the UPMC Food-101 and Hateful Memes datasets. $\mu_{w/o}$ and $\mu_{w/}$ represent the average cosine similarity values w/o and w/ projection, respectively.

Analysis on Inter-task Gradient Projection. To assess the impact of inter-task gradient projection, we visualize the cosine similarity between the gradients of KL divergence (g_{kl}) and the target task ($g_{\mathcal{T}}$) throughout the entire training process using histograms, as shown in Figure 7. Without gradient projection, negative similarity between gradients is prevalent throughout training, resulting in imbalanced updates to the target task. Conversely, BALGRAD, incorporating inter-task gradient projection, shows a positive mean cosine similarity between gradients, indicating fewer conflicts during training. This suggests that the gradients for the target task are more balanced between the two modalities, leading to more balanced convergence. This reduction in conflicts narrows the performance gap between image and text modalities, mitigating over-reliance on any specific modality, aligning with our analysis in Section 3.1.

To further quantify this effect, we conduct an ablation study measuring the frequency of conflicting gradients with and without projection across three datasets, as shown in Table 3. The fraction indicates the percentage of gradient conflicts that occur between the gradients of KL divergence (g_{kl}) and the target task ($g_{\mathcal{T}}$) throughout the entire training process. The results demonstrate that there is a high incidence of conflicting gradients across all datasets without projection. In contrast, the use of projection significantly reduces gradient conflicts, especially in datasets with dominant modality bias, such as UPMC Food-101 and Hateful Memes.

	UPMC Food-101		Hateful Memes		MM-IMDb	
	Fraction↓	Δ_{Gap}↓	Fraction↓	Δ_{Gap}↓	Fraction↓	Δ_{Gap}↓
w/o Projection	0.66	43.27	0.78	10.21	0.28	4.21
w/ Projection	**0.36**	**39.54**	**0.32**	**8.28**	**0.26**	**4.04**

Table 3: Ablative results show the fraction of conflicting gradients and Δ_{Gap} on the UPMC Food-101 and Hateful Memes datasets, comparing scenarios without inter-task gradient projection ("w/o Projection") and with standard BALGRAD ("w/ Projection").

5 Conclusion

In this paper, we addressed the challenge of dominant modality bias, where a VL model disproportionately relies on one modality, undermining the contributions of others. Our analysis shows that unaligned gradients and differences in gradient *magnitudes* hinder balanced loss convergence. Based on these findings, BALGRAD mitigates this bias by incorporating inter-modality gradient reweighting, which adjusts the KL divergence gradient based on each modality's contribution, and inter-task gradient projection to align task *directions* non-conflictingly. Experiments on UPMC Food-101, Hateful Memes, and MM-IMDb datasets demonstrate that BALGRAD effectively reduces dominant modality bias, enhances model robustness, and improves accuracy. These results highlight the potential for more stable and balanced training in VL models, paving the way for future advancements.

Limitation

While BALGRAD has shown efficacy in mitigating dominant modality bias in VL models, extending this approach to multimodal models with more than two modalities presents additional challenges. When dealing with three or more modalities, the training cost rapidly increases due to the need to consider the relationships between the gradients of each pair of modalities. This increased complexity in gradient management makes the balancing process more computationally intensive and difficult to maintain effectively. Thus, while BALGRAD is effective in bi-modal settings, its application in multimodal scenarios requires further refinement to manage the higher computational demands and ensure balanced performance across all modalities.

Acknowledgment

This research was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01341, Artificial Intelligence Graduate School Program, Chung-Ang University). This research was also supported by the MSIT (Ministry of Science and ICT), Korea , under the Graduate School of Metaverse Convergence support program (IITP-2025-RS-2024-00418847) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Piush Aggarwal, Jawar Mehrabanian, Weigang Huang, Özge Alacam, and Torsten Zesch. 2024. Text or image? what is more important in cross-domain generalization capabilities of hate meme detection models? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 104–117.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. In *Proceedings of the International Conference on Learning Representations: Workshop Track*.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 233–242.

Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. 2022. Multimodality for nlp-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. 2023. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *Proceedings of the NeurIPS 2014 Deep Learning and Representation Learning Workshop*.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989.

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

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. Mm-reasoner: A multimodal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, pages 2611–2624.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900. PMLR.

Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. 2023. Biasadv: Bias-adversarial augmentation for model debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3832–3841.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *Proceedings of the International Conference on Learning Representations*.

Andrei Manolache, Florin Brad, and Elena Burceanu. 2021. Date: Detecting anomalies in text via self-supervision of transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 267–277.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2020. A multimodal dataset of images and text to study abusive language. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769. CEUR-WS. org.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv preprint arXiv:2303.09892*.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Debiasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, pages 20673–20684.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 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.

Mary Phuong and Christoph Lampert. 2019. Towards understanding knowledge distillation. In *Proceedings of the International Conference on Machine Learning*, pages 5142–5151.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Guangyuan Shi, Qimai Li, Wenlong Zhang, Jiaxin Chen, and Xiao-Ming Wu. 2023. Recon: Reducing conflicting gradients from the root for multi-task learning. In *Proceedings of the International Conference on Learning Representations*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. 2015. Recipe recognition with large multimodal food dataset. In *Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops*, pages 1–6.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11293–11302.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, California Institute of Technology*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. 2023. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2776–2784.

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

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. 2024. Quantifying and enhancing multi-modal robustness with modality preference. In *Proceedings of the International Conference on Learning Representations*.

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

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 5824–5836.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2023. Hype: Better pre-trained language model fine-tuning with hidden representation perturbation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3246–3264.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328.

A Appendix of Propositions

A.1 Proof of Proposition 1

Proposition 1. (Gradient Effect on Change of Loss) Let the parameters θ_v, θ_l, and $\theta_\mathcal{T}$ of a multimodal model be updated with gradients g_v, g_l, and $g_\mathcal{T}$ using a sufficiently small step size $\lambda > 0$, resulting in updated parameters $\hat{\theta}_v, \hat{\theta}_l$, and $\hat{\theta}_\mathcal{T}$. Then the change in the loss function satisfies

$$\begin{aligned}\triangle\mathcal{L} \;=\; & -2\,\lambda\,\left(g_\mathcal{T}^v \cdot g_\mathcal{T}^l\right) \\ & -\,\lambda \sum_{i\in\{v,l,\mathcal{T}\}} \left(g_i \cdot g_i \;+\; g_\mathcal{T}^i \cdot g_\mathcal{T}^i\right) \;+\; O(\lambda^2),\end{aligned} \tag{7}$$

where the cross term $-2\,\lambda\,(g_\mathcal{T}^v) \cdot (g_\mathcal{T}^l)$ captures the interaction between the visual and language gradients and the magnitudes and directions of each gradient $g_\mathcal{T}^v$ and $g_\mathcal{T}^l$ governs how much the overall loss is reduced.

Proof of Proposition 1.

Proof. Let the $\theta_v, \theta_l, \theta_\mathcal{T}$ be updated in the direction of negative gradients $g_v, g_l, g_\mathcal{T}$ with step size $\lambda > 0$. Then the updated $\hat{\theta}_v, \hat{\theta}_l, \hat{\theta}_\mathcal{T}$ are $\theta_v - \lambda g_v, \theta_l - \lambda g_l, \theta_\mathcal{T} - \lambda g_\mathcal{T}$. In that case, the change in the loss function with updated parameters is

$$\triangle\mathcal{L} = \mathcal{L}(\hat{\theta}_v, \hat{\theta}_l, \hat{\theta}_\mathcal{T}) - \mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T})$$

By the first-order taylor expansion with a point $(\theta_v, \theta_l, \theta_\mathcal{T})$,

$$\begin{aligned}
&\mathcal{L}(\hat{\theta}_v, \hat{\theta}_l, \hat{\theta}_\mathcal{T}) - \mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T}) \\
&= \mathcal{L}(\theta_v - \lambda g_v, \theta_l - \lambda g_l, \theta_\mathcal{T} - \lambda g_\mathcal{T}) \\
&\quad - \mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T}) \\
&= \mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T}) \\
&\quad + (\theta_v - \lambda g_v - \theta_v)^T g_v (\theta_l - \lambda g_l - \theta_l)^T g_l \\
&\quad + (\theta_\mathcal{T} - \lambda g_\mathcal{T} - \theta_\mathcal{T})^T g_\mathcal{T} \\
&\quad - \mathcal{L}(\theta_v, \theta_l, \theta_\mathcal{T}) + + O(\lambda^2) \\
&= -\lambda(g_v^T \cdot g_v + g_l^T \cdot g_l) \\
&\quad - \lambda(g_\mathcal{T}^l + g_\mathcal{T}^v)^T (g_\mathcal{T}^l + g_\mathcal{T}^v) \\
&\quad + O(\lambda^2) \\
&= -2\lambda g_\mathcal{T}^v \cdot g_\mathcal{T}^l - \lambda \sum_{i\in\{v,l,\mathcal{T}\}} (g_i \cdot g_i + g_\mathcal{T}^i \cdot g_\mathcal{T}^i) \\
&\quad + O(\lambda^2)
\end{aligned}$$

□

Influence of Fusion Methods. The term $\left(g_\mathcal{T}^v\right)^\top \left(g_\mathcal{T}^l\right)$ captures how the visual and language gradients interact within the classifier parameters $\theta_\mathcal{T}$. Different fusion methods yield different dependencies of $g_\mathcal{T}^v$ and $g_\mathcal{T}^l$ on v and l:

- **Addition:** The fused input is $x = v + l$. Because v and l are merged by simple addition, their representations feed directly into the same part of the classifier. Consequently, $\left(g_\mathcal{T}^v\right)^\top \left(g_\mathcal{T}^l\right)$ often remains significant due to the shared pathway.
- **Concatenation:** The fused input is $x = [v; l]$. Each modality is placed in distinct segments of the classifier's input vector, reducing direct interactions. As a result, $g_\mathcal{T}^v$ and $g_\mathcal{T}^l$ may be more independent, potentially lowering the cross term $\left(g_\mathcal{T}^v\right)^\top \left(g_\mathcal{T}^l\right)$.
- **Attention:** The fused input is $x = \mathrm{Attention}(v, l)$. This method can create strong interdependence between v and l within $\theta_\mathcal{T}$. Hence, it $\left(g_\mathcal{T}^v\right)^\top \left(g_\mathcal{T}^l\right)$ can become highly influential since changes v affect l and vice versa through the attention mechanism.

Hence, the sign and magnitude of the cross term reflect how strongly the parameters for the two modalities are tied together under each fusion strategy.

A.2 Proof of proposition 2

Proposition 2. (Gradient Conflicts on Loss Reduction with KL Loss) Let $\mathcal{G}^{\tau} = \{ g_v, g_l, g_{\tau}\}$ and $\mathcal{G}^{kl} = \{ g_v^{kl}, g_l^{kl}, 0\}$ be the gradients from a target loss $\mathcal{L}_{\tau}$ and a KL loss $\mathcal{L}_{kl}$, respectively, with parameters $\theta = [\theta_v, \theta_l, \theta_{\tau}]^{\top}$. Assume the parameters are updated by gradient descent with a small step size $\lambda > 0$: $\theta_v' = \theta_v - \lambda \left(g_v + g_v^{kl}\right), \quad \theta_l' = \theta_l - \lambda \left(g_l + g_l^{kl}\right), \quad \theta_{\tau}' = \theta_{\tau} - \lambda\, g_{\tau}$. Then, for the combined loss $\mathcal{L} = \mathcal{L}_{\tau} + \mathcal{L}_{kl}$, the change in the loss is

$$\begin{aligned}\Delta\mathcal{L} &= \mathcal{L}\big(\theta'\big) - \mathcal{L}\big(\theta\big) \\ &= -\lambda\Big(\|\mathcal{G}^{\tau}\|^2 + \|\mathcal{G}^{kl}\|^2 + 2\,(\mathcal{G}^{\tau})^{\top}\mathcal{G}^{kl}\Big) \\ &\quad + O(\lambda^2).\end{aligned} \tag{8}$$

In particular, if $(\mathcal{G}^{\tau})^{\top}\mathcal{G}^{kl} < 0$, the gradients from the target and KL losses *conflict*, reducing the effective loss reduction.

Proof of Proposition 2.

Proof. Because $\mathcal{L} = \mathcal{L}_{\tau} + \mathcal{L}_{kl}$, its gradient is

$$\nabla_{\theta}\mathcal{L} = \mathcal{G}^{\tau} + \mathcal{G}^{kl}.$$

Under a small step size λ, a first-order Taylor expansion about θ gives $\Delta\mathcal{L} \approx -\lambda\,\|\mathcal{G}^{\tau} + \mathcal{G}^{kl}\|^2$. Since $\mathcal{G}^{\tau} = \{g_v^{\tau}, g_l^{\tau}, g_{\tau}^{\tau}\}$ and $\mathcal{G}^{kl} = \{g_v^{kl}, g_l^{kl}, 0\}$, the relevant parameters are updated as:

$$\begin{aligned}\theta_v' &= \theta_v - \lambda\,(g_v^{\tau} + g_v^{kl}) \\ \theta_l' &= \theta_l - \lambda\,(g_l^{\tau} + g_l^{kl}) \\ \theta_{\tau}' &= \theta_{\tau} - \lambda\, g_{\tau}^{\tau}.\end{aligned}$$

By decomposing norm:

$$\begin{aligned}\Delta\mathcal{L} = &-\lambda\Big(\|g_v^{\tau}\|^2 + \|g_v^{kl}\|^2 + 2\,g_v^{\tau})^{\top}(g_v^{kl}) \\ &+ \|g_l^{\tau}\|^2 + \|g_l^{kl}\|^2 + 2\,g_l^{\tau})^{\top}(g_l^{kl}) + \|g_{\tau}^{\tau}\|^2\Big) \\ &+ O(\lambda^2)\end{aligned}$$

Hence, if either $(g_v^{\tau})^{\top}(g_v^{kl}) < 0$ or $(g_l^{\tau})^{\top}(g_l^{kl}) < 0$, the negative cross-term reduces the effective loss decrease for that modality. □

B Further Implementation Details

B.1 Dataset and Evaluation Metrics

UPMC Food-101 (Wang et al., 2015) is a food classification dataset with 101 categories and 90,840 image-text pairs, involving the classification of food items using both images and textual recipe descriptions; to create a validation split, we extracted 5,000 samples from the training set (Kiela et al., 2019), as the dataset only provides training and testing sets.

Hateful Memes (Kiela et al., 2020) is designed to detect hate speech by combining image and text modalities, comprising 8,500 training samples, 1,000 validation samples, and 500 test samples.

MM-IMDb (Arevalo et al., 2017) is a multi-label movie genre classification dataset that incorporates poster images and plot descriptions, containing 23 genre tags with 15,552 training samples, 2,608 validation samples, and 7,799 test samples.

We utilize classification accuracy, AUROC, and F1-Macro as evaluation metrics for the UPMC Food-101, Hateful Memes, and MM-IMDb datasets, respectively.

B.2 Architecture and Training Scheme

In all comparative experiments, we employ ViT (Dosovitskiy et al., 2021) and BERT (Devlin et al., 2019) as image and text encoders, respectively. We adopt a late concatenation architecture where the embeddings from each modality are concatenated to make the final prediction. We employ linear probing as our fine-tuning strategy, which freezes all the encoder parameters and trains only the embedding and classifier layers.

We adopt this modular architecture and fine-tuning scheme for several key reasons: First, the modular design of BALGRAD allows it to extend to various encoders, easily accommodating different architectures. This flexibility is crucial in real-world scenarios where resources are often constrained. Our structure supports a range of scalable encoder configurations, ensuring adaptability to different resource availability and application requirements. Additionally, in some cases, data access is restricted due to privacy concerns, necessitates the use of pre-extracted features (Cheplygina et al., 2019; Kruk et al., 2019; Menini et al., 2020), making the application of early fusion-based large VLMs (Liu et al., 2024) impractical. Also, to focus improvements on BALGRAD's gradient reweighting and projection, we adopt linear probing as a

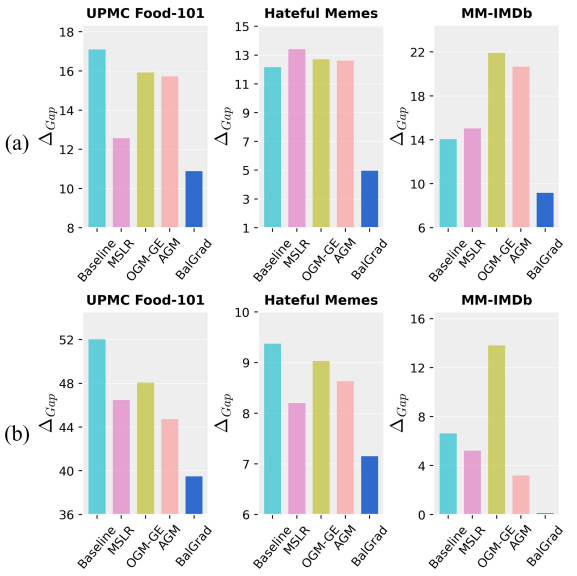


Figure 8: Bar plots illustrating the performance of existing methods and BALGRAD with different fusion mechanisms: **(a)** addition and **(b)** attention, evaluated on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets. Each bar indicates $\Delta_{Gap}(\%)$, which quantifies the performance variation between missing image and missing text conditions.

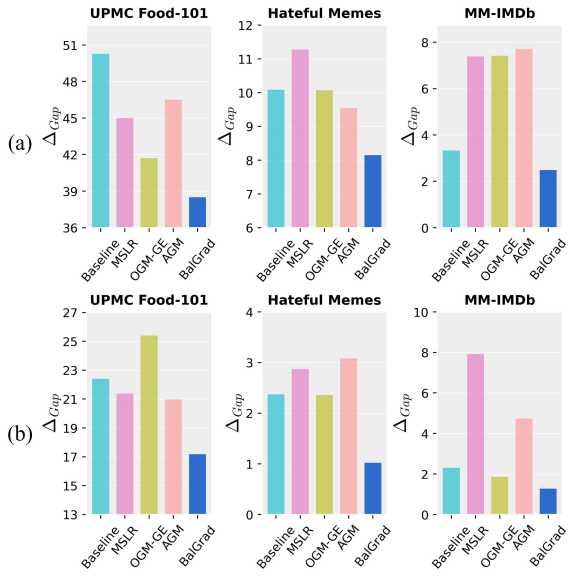


Figure 9: Bar plots presenting the performance comparison between existing methods and BALGRAD across different backbone models: **(a)** ResNet and DistilBERT, and **(b)** CLIP, on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets. Each bar represents $\Delta_{Gap}(\%)$, measuring the performance discrepancy under missing image and missing text conditions.

fine-tuning strategy, ensuring that the gains were not merely due to the encoders' inherent capabilities but to our method's effectiveness.

As a baseline, we adopt a standard linear probing approach and compare our proposed method against existing methods designed to balance modalities in VL models, specifically MSLR (Yao and Mihalcea, 2022), OGM-GE (Peng et al., 2022), and AGM (Li et al., 2023).

B.3 Implementation Details

We use `vit-base` and `bert-base-uncased` checkpoints as the image and text encoders, respectively, loading them from Transformers library (Wolf et al., 2020). The embeddings extracted from each encoder have a dimensionality of 768, and we concatenate these embeddings to form a 1568-dimensional vector, which is then passed to a final classifier. We resize all images to 224×224 and apply a random horizontal flip for augmentation. For text, the maximum sequence lengths are set to 1024 for MM-IMDb, 512 for UPMC Food-101, and 128 for Hateful Memes. We use the Adam optimizer with a momentum of 0.9 for all experiments, training for 20 epochs with a batch size of 128.

C Additional Experimental Results

C.1 Experimental Results on Different Fusion Mechanisms

The way embeddings from different modalities are fused can significantly impact a model's ability to capture and leverage cross-modal interactions. We conducted experiments on different fusion strategies in the baseline and BALGRAD, specifically exploring element-wise addition and attention-based fusion mechanisms following previous work (Kumar and Nandakumar, 2022). We tested these mechanisms on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets, evaluating the Δ_{Gap} in performance under conditions where either the image or text modality was missing. Results for addition and attention are presented in Figure 8. Across all datasets, BALGRAD demonstrated the smallest Δ_{Gap} with both fusion mechanisms, effectively mitigating dominant modality bias. This confirms that BALGRAD effectively captures and leverages cross-modal interactions across different fusion mechanisms.

C.2 Experimental Results on Different Backbone Models

We conduct extensive experiments across diverse backbone models, underscoring its consistent per-

formance and adaptability to varying architectures and computational resources. Specifically, we employ lower-capacity models—ResNet-50 (He et al., 2016) for the image encoder and DistilBERT (Sanh et al., 2019) for the text encoder—to assess robustness concerning model size. Additionally, we leverage the widely-used multimodal pretrained VLM, CLIP (Radford et al., 2021), for further evaluation due to its strong ability to seamlessly integrate visual and textual information, providing a rigorous test for BALGRAD. Experiments were carried out on the UPMC Food-101, Hateful Memes, and MM-IMDb datasets, assessing the performance gap under conditions where either the image or text modality was missing. Results for ResNet-DistilBERT and CLIP are presented in Figure 9. Across all datasets, BALGRAD consistently exhibited the smallest Δ_{Gap}, effectively balancing the contributions between modalities.

Intriguingly, while earlier experiments using ViT and BERT encoders on the MM-IMDb dataset showed no over-reliance on a specific modality, our additional studies reveal that conventional methods tend to heavily rely on the text modality, when employing ResNet and DistilBERT. These findings indicate that such bias is influenced not only by the task but also by the choice of backbone model. Our comprehensive experiments affirm that BALGRAD effectively mitigates bias irrespective of the backbone model employed, showcasing its superior scalability.

C.3 Experimental Results with Additional Datasets

To validate the generalizability of BALGRAD, we conduct experiments on two additional datasets: Memotion (Mishra et al., 2023) and CUB-200-2011 (Welinder et al., 2010). The Memotion dataset, used for classifying the humor level of meme images based on their descriptions, includes annotations such as "not funny", "funny", "very funny", and "hilarious". The CUB-200-2011 dataset is a fine-grained bird classification dataset, requiring the categorization of 200 bird species based on images and descriptions. We evaluate each dataset using weighted F1 score and classification accuracy.

The results for the Memotion dataset, presented in Table 4, show that when the text modality is missing, performance drops significantly more than when the image modality is missing, indicating a bias toward the text modality. BALGRAD not only achieves the highest performance with the image modality alone but also excels in the *Avg.* and Δ_{Gap} metrics, demonstrating effective modality balance.

Modality		Memotion				
		Baseline	MSLR	OGM-GE	AGM	BALGRAD
Full		70.55	70.36	70.28	**71.12**	<u>70.77</u>
Missing	Image	58.34	59.24	**59.66**	59.54	59.48
	Text	49.29	51.32	50.38	<u>51.44</u>	**52.78**
	Avg.↑	53.82	55.28	55.02	<u>55.49</u>	**56.13**
	Δ_{Gap}↓	4.53	<u>3.96</u>	4.64	4.05	**3.35**

Table 4: The experimental result of BALGRAD on the Memotion dataset. The best result in each test dataset is boldfaced, and the second best is presented with underlining. "*Avg.*" represents the average performance under conditions where one of the modality is missing, while "$\Delta_{Gap}(\%)$" indicates the performance difference.

Modality		CUB-200-2011				
		Baseline	MSLR	OGM-GE	AGM	BALGRAD
Full		74.71	72.12	75.15	**76.28**	<u>75.84</u>
Missing	Image	37.38	40.21	39.49	<u>41.24</u>	**45.47**
	Text	61.24	60.20	59.14	<u>61.42</u>	**62.72**
	Avg.↑	49.31	50.21	49.32	<u>51.33</u>	**54.10**
	Δ_{Gap}↓	11.93	9.99	<u>9.83</u>	10.09	**8.63**

Table 5: The results of BALGRAD on the CUB-200-2011 dataset are presented. The highest performance in each test dataset is shown in bold, with the second-highest underlined. "*Avg.*" reflects the average performance when one modality is absent, and "$\Delta_{Gap}(\%)$" denotes the performance difference.

As shown in Table 5, the CUB-200-2011 dataset exhibits a strong reliance on the image modality. BALGRAD outperforms AGM by more than 4%p in accuracy when the image modality is missing and achieves the smallest Δ_{Gap} at 8.63%, demonstrating its superiority in handling fine-grained classification tasks even under challenging conditions.