Balancing Multimodal Domain Generalization via Gradient Modulation and Projection

Hongzhao Li[1], Guohao Shen[1], Shupan Li[1,2,3*], Mingliang Xu[1,2,3,4*], Muhammad Haris Khan[5]

[1]School of Computer and Artificial Intelligence, Zhengzhou University
[2]Engineering Research Center of Intelligent Swarm Systems, Ministry of Education
[3]National Supercomputing Center in Zhengzhou
[4]Shandong Bosuan Zhixin Information Technology Co., Ltd
[5]Mohamed Bin Zayed University of Artificial Intelligence
{lihongzhao, sgh_2024}@gs.zzu.edu.cn, {iespli, iexumingliang}@zzu.edu.cn, muhammad.haris@mbzuai.ac.ae

Abstract

Multimodal Domain Generalization (MMDG) leverages the complementary strengths of multiple modalities to enhance model generalization on unseen domains. A central challenge in multimodal learning is optimization imbalance, where modalities converge at different speeds during training. This imbalance leads to unequal gradient contributions, allowing some modalities to dominate the learning process while others lag behind. Existing balancing strategies typically regulate each modality's gradient contribution based on its classification performance on the source domain to alleviate this issue. However, relying solely on source-domain accuracy neglects a key insight in MMDG: modalities that excel on the source domain may generalize poorly to unseen domains, limiting cross-domain gains. To overcome this limitation, we propose Gradient Modulation Projection (GMP), a unified strategy that promotes balanced optimization in MMDG. GMP first decouples gradients associated with classification and domain-invariance objectives. It then modulates each modality's gradient based on semantic and domain confidence. Moreover, GMP dynamically adjusts gradient projections by tracking the relative strength of each task, mitigating conflicts between classification and domain-invariant learning within modality-specific encoders. Extensive experiments demonstrate that GMP achieves state-of-the-art performance and integrates flexibly with diverse MMDG methods, significantly improving generalization across multiple benchmarks.

Introduction

Multimodal Domain Generalization (MMDG) is an emerging research area focused on training models that generalize to unseen domains by leveraging multiple modalities, such as video and audio (Dong et al. 2025a,b). Unlike traditional unimodal approaches (Munir et al. 2023; Khan, Shaaban, and Khan 2024; Galappaththige et al. 2024; Galappaththige, Kuruppu, and Khan 2024), MMDG exploits the complementary strengths of different modalities to boost both robustness and generalization. This advantage is especially important in real-world applications like cross-environment action recognition, audiovisual event detection, and multimodal surveillance, where test data often originate from novel domains, such as different environments, devices, or conditions (Zhou et al. 2022). In such cases, effective generalization hinges on learning domain invariant features without compromising discriminative power (Wang et al. 2022).

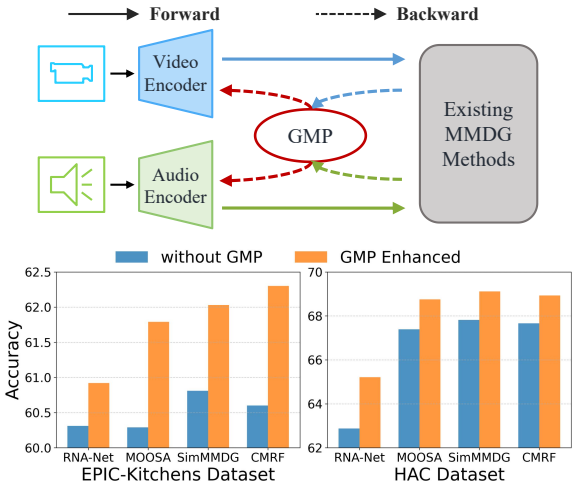

Figure 1: Performance of state of the art MMDG methods, RNA-Net (Planamente et al. 2024), MOOSA (Dong, Chatzi, and Fink 2024), SimMMDG (Dong et al. 2023), and CMRF (Fan et al. 2024), evaluated with and without our proposed GMP strategy.

A common challenge in multimodal learning is the optimization imbalance between modalities (Wei et al. 2024). During training, different modalities often converge at varying rates, leading to uneven gradient updates. When one modality dominates, it can hinder learning in others, disrupting balanced training dynamics. This imbalance reduces the effectiveness of multimodal learning, as the model becomes overly dependent on the dominant modality while underutilizing the rest (Peng et al. 2022). In MMDG, this problem is especially evident. As shown in Table 1, the performance of each unimodal branch in a typical video-audio MMDG model is significantly worse than when trained indepen-

*Corresponding authors.

dently. This indicates that current MMDG training strategies fail to capitalize on the strengths of each modality, limiting the model's ability to generalize across unseen domains.

The imbalance issue in MMDG is more intricate than in standard multimodal learning due to its dual objectives: accurate classification and domain invariant generalization. Traditional imbalance mitigation methods focus mainly on improving classification within the source domain. However, this neglects a key insight: a modality that excels in source domain classification may still fail to learn domain invariant features effectively. As a result, balancing solely based on classification performance often has limited impact on generalization to target domains. This limitation is empirically demonstrated in Table 1, where traditional methods show improved source domain performance but limited generalization to unseen domains.

To address these challenges, we propose a unified optimization strategy for MMDG, termed Gradient Modulation Projection (GMP) (Fig. 2), designed to fully exploit the domain generalization capabilities of each modality. GMP integrates two key components. The first component, Inter-Modality Gradient Decoupled Modulation (IGDM), differs from conventional methods by separating gradients for classification and domain invariance tasks. This decoupling allows the model to independently assess and optimize each modality's contribution to both objectives. IGDM employs a dual modulation approach guided by two confidence metrics: semantic confidence for classification and domain confidence for generalization. These metrics enable dynamic, task specific gradient modulation based on each modality's strengths. The second component, Conflict-Adaptive Gradient Projection (CAGP), addresses gradient conflicts within modality specific encoders. Through theoretical analysis, we show that classification and domain invariance gradients often point in opposing directions, impeding effective optimization. CAGP mitigates this by continuously monitoring task intensities during training and dynamically adjusting gradient projections. When conflicts arise, the relatively weaker task gradient is preserved, while the stronger one is projected orthogonally to reduce interference.

As shown in Table 1, existing balancing strategies such as OGM-GE (Wang, Tran, and Feiszli 2020) and Grad-Blending (Peng et al. 2022) achieve notable improvements on the source domain compared to the baseline model, but yield only marginal gains on the target domain (e.g., +0.65% and +0.43% in the video-audio task, respectively). In contrast, our proposed GMP achieves a similar improvement on the source domain while delivering a substantially higher gain on the target domain (+2.30%). These results demonstrate that GMP maintains strong source domain accuracy without sacrificing generalization, promoting more balanced multimodal learning.

Our contributions are summarized as follows:

- We analyze why traditional multimodal learning balancing approaches fail to resolve optimization imbalance in MMDG. Empirically, we show that they often prioritize source domain classification, while neglecting generalization to unseen domains.

Domain	Method	Video	Audio	Video-Audio
Source	Uni-video	74.55	-	-
	Uni-audio	-	52.35	-
	Base	70.28	48.63	76.31
	OGM-GE	72.66	50.33	77.69
	Grad-Blending	72.31	50.62	78.03
	GMP(Ours)	72.69	50.52	78.05
Target	Uni-video	54.98	-	-
	Uni-audio	-	38.86	-
	Base	48.86	34.15	55.06
	OGM-GE	49.75	35.33	55.71
	Grad-Blending	50.03	35.26	55.49
	GMP(Ours)	**52.33**	**35.88**	**57.36**

Table 1: Performance comparison on source and target sets using the EPIC-Kitchens dataset (Kay et al. 2017). Source results are averaged over validation sets (of the seen domains), and target results are averaged over test sets from the corresponding unseen domains. Uni-video and Uni-audio denote independently trained single-modality models.

- We introduce GMP, a unified strategy that adjusts classification and domain invariance gradients through IGDM and resolves inter-task conflicts using CAGP.
- To the best of our knowledge, this is the first work to examine and address MMDG from an optimization perspective, offering a new path for MMDG task.
- Extensive experiments demonstrate that our unified GMP strategy not only achieves state of the art performance but also integrates smoothly with diverse MMDG methods (Fig. 1), showcasing strong versatility.

Related Work

Multimodal Domain Generalization

Unimodal domain generalization (DG) has been extensively studied, providing a solid foundation for understanding model generalization across domains. In contrast, MMDG remains in its early stages, with recent efforts focusing on multimodal alignment and representation learning (Li et al. 2025b). RNA-Net (Planamente et al. 2022) introduces a relative canonical alignment loss to balance audio and video features, tackling the alignment challenges of heterogeneous modalities. SimMMDG (Dong et al. 2023) decouples modalities to better capture semantic structures and enhance domain invariant representations. MOOSA (Dong, Chatzi, and Fink 2024) leverages an auxiliary pretext task to promote cross modal relationships, thereby improving feature representations. Similarly, CMRF (Fan et al. 2024) flattens cross modal representations to strengthen feature alignment and consistency. While these methods contribute to MMDG, they primarily emphasize domain invariant learning and overlook optimization dynamics between modalities, an underexplored yet crucial issue, as intermodal imbalance during training can hinder generalization.

Imbalanced Multimodal Learning

Multimodal models often favor modalities that are easier to learn, limiting the contribution of more complex ones (Ma, Liu, and Cheng 2024; Li et al. 2025e,d,c,a). Despite access to diverse multimodal data, performance improvements remain constrained by inherent modality disparities, such as varying convergence speeds. To address this, several approaches seek to balance learning across modalities. GradBlending (Wang, Tran, and Feiszli 2020) exploits overfitting tendencies to optimize modality combinations. OGM-GE (Peng et al. 2022) builds on this by adaptively controlling gradients to maintain intermodal balance. AGM (Li et al. 2023) uses Shapley values to adjust gradients based on single modal responses, encouraging equitable learning. CGGM (Guo et al. 2024) regulates gradient magnitude and direction using classifier guided signals to address imbalance. DRB (Wei et al. 2024) estimates each modality's learning state through the separability of its single modal representation, then softly reinitializes encoders to prevent overfitting to less informative inputs. BALGRAD (Kwon et al. 2025) further reduces intermodal differences through alignment to encourage balanced learning. Although these methods enhance performance on seen domains, they often underperform in MMDG settings due to a key limitation: they fail to account for each modality's domain generalization capability, reducing robustness on unseen domains.

Preliminaries

Training Setting

The training dataset $\mathcal{T} = \{x_i, y_i, d_i\}_{i=1}^{N}$ consists of samples from C source domains, where $d_i \in \{1, \ldots, C\}$ denotes the domain label and $y_i \in \{1, \ldots, Y\}$ the class label for each sample i. Each input $x_i = (x_i^v, x_i^a)$ includes video ($m = v$) and audio ($m = a$) modalities. The objective is to learn class-informative, domain invariant features that generalize to unseen domains by balancing two goals: (1) classification, which promotes class-informative features, and (2) domain invariance, which mitigates domain specific biases. To this end, we adopt a combined loss function used in unimodal domain generalization (Li et al. 2018) and adapt it to the multimodal case, where the classification loss L_c and domain adversarial loss L_d are computed via a classifier and a domain discriminator, respectively:

$$L = L_c + \lambda L_d, \tag{1}$$

where λ controls the trade-off between the two objectives. At training step t, let θ_t^m denote the parameters for modality $m \in \{v, a\}$. The classification and domain gradients with respect to θ_t^m are $\nabla_{\theta_t^m} L_c$ and $\nabla_{\theta_t^m} L_d$, respectively. Incorporating trade-off λ, we define the gradients for modality m at step t:

$$g_c^m = \nabla_{\theta_t^m} L_c, \quad g_d^m = (-\lambda) \cdot \nabla_{\theta_t^m} L_d. \tag{2}$$

where the negative sign in g_d^m implements gradient reversal. The parameter update:

$$\begin{aligned} \theta_{t+1}^m &= \theta_t^m - \eta \left(\nabla_{\theta^m} L_c - \lambda \nabla_{\theta^m} L_d \right) \\ &= \theta_t^m - \eta \left(g_c^m + g_d^m \right) \\ &= \theta_t^m - \eta G_t^m, \end{aligned} \tag{3}$$

where G_t^m denotes the total gradient applied to modality m at step t. We now analyze the unbalanced forms in MMDG: inter-modality imbalance and inter-task imbalance.

Inter-Modality Imbalance

In multimodal learning, persistent discrepancies in gradient magnitudes across modalities can result in one modality dominating the optimization process, thereby diminishing contributions of others (Peng et al. 2022). Specifically, if $|g_c^v| \gg |g_c^a|$, the video modality exerts a disproportionately strong influence. This dominance is quantified by the ratio $r_{v,a}(t) = ||g_c^v|| / ||g_c^a|| \gg 1$ indicating that the video gradients dominate the updates at step t. Furthermore, this imbalance accumulates significantly over T steps as $R_{v,a}(T) = (\sum_{t=1}^{T} \|g_c^v\|) / (\sum_{t=1}^{T} \|g_c^a\|)$, which leads to chronic under optimization of weaker modalities like audio.

However, MMDG introduces a critical nuance in which conventional gradient modulation strategies that balance modalities based solely on classification performance (reflected in $|g_c^m|$) prove fundamentally inadequate. This inadequacy arises for several reasons. There is a generalization critical asymmetry where a modality with strong source domain discriminative power may exhibit weak domain invariance, whereas a seemingly weaker modality could excel at capturing domain agnostic features essential for unseen domains. Therefore, simply balancing via $R_{v,a}(T)$, which ignores invariance quality, risks suppressing modalities that are vital for cross domain generalization. As shown in Table 1, traditional balancing approaches often fail on MMDG benchmarks because they favor source domain accuracy over robust generalization.

Inter-Task Conflicts

Another challenge arises when gradients from different tasks, such as classification and domain adversarial training, conflict with each other and exhibit negative cosine similarity (Yu et al. 2020). In such cases, a dominant gradient may overshadow others, degrading the performance of less represented tasks. In DG tasks, this issue often arises when the classification loss gradient g_c^m and the domain adversarial loss gradient g_d^m point in opposing directions.

Let $g_c^m = \{g_c^v, g_c^a\}$ and $g_d^m = \{g_d^v, g_d^a\}$ represent the gradients of the classification loss L_c and the domain adversarial loss L_d, respectively, where $\theta = [\theta^v, \theta^a]$ are the modality specific parameters for video and audio. Assuming updates are performed using gradient descent, the parameter updates become:

$$\theta_{t+1}^v = \theta_t^v - \eta \left(g_c^v + g_d^v \right), \quad \theta_{t+1}^a = \theta_t^a - \eta \left(g_c^a + g_d^a \right).$$

For the combined loss $L = L_c + \lambda L_d$, the first order approximation of the change in total loss after a single update step is:

$$\begin{aligned} \Delta L &= L(\theta_{t+1}) - L(\theta_t) \\ &= -\eta \left(||g_c^m||^2 + ||g_d^m||^2 + 2 {g_c^m}^\top g_d^m \right) + \mathcal{O}(\eta^2). \end{aligned} \tag{4}$$

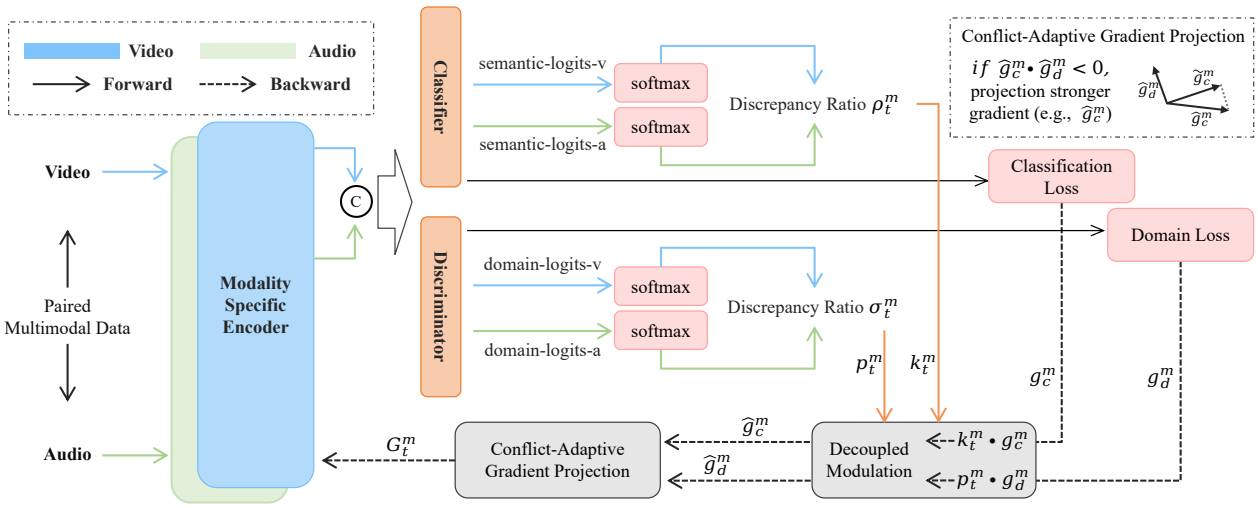

Figure 2: The overall training strategy of our proposed GMP.

If $g_c^{m\top} g_d^m < 0$, the classification and domain adversarial gradients are in conflict, which misalignment can impede convergence and deteriorate multi task optimization.

In MMDG, such conflicts are often modality specific. For instance, video related gradients may show stronger conflict between classification and domain objectives, while audio gradients remain more aligned. Consequently, applying a uniform conflict resolution strategy across modalities fails to capture these modality specific dynamics. As demonstrated in Table 4, traditional conflict mitigation strategies underperform, ultimately limiting generalization to unseen domains.

Method

Inter-Modality Gradient Decoupled Modulation

To address inter-modality imbalance, we propose an Inter-Modality Gradient Decoupled Modulation (IGDM) strategy for MMDG. Unlike traditional unified gradient balancing methods that mainly consider overall modality classification ability, IGDM introduces a decoupled modulation strategy that adjusts each modality's contribution independently for classification and domain discrimination tasks. This fine grained control directly tackles the shortcomings of classification gradient only balancing strategies.

We consider video and audio modalities, each processed by a modality specific encoder $\varphi^m(\theta^m, \cdot)$ that extracts features from the input. These features are concatenated and fed into a classifier f_c and a domain discriminator f_d:

$$\begin{aligned} f_c(x_i) &= \sum_{m=1}^{M} W^m \varphi^m(\theta^m, x_i^m) + b_y, \\ f_d(x_i) &= \sum_{m=1}^{M} D^m \varphi^m(\theta^m, x_i^m) + b_d, \end{aligned} \tag{5}$$

where M is the number of modalities, W^m and D^m are the respective weights for classification and domain discrimination, and b_y, b_d are the bias terms.

As established in prior work (Peng et al. 2022), modalities with higher prediction confidence tend to contribute less to the joint gradient through their respective weights. This results in dominant modalities suppressing gradient updates for weaker ones, leading to biased optimization. To quantify individual modality contributions and counteract this effect, we introduce two distinct confidence metrics. Semantic Confidence measures classification certainty; higher values indicate stronger discriminative power. Domain Confidence measures domain confusion; lower values indicate stronger domain invariant features.

For each modality m and sample i, we compute semantic confidence q_i^m using its contribution from the classifier:

$$q_i^m = \sum_{k=1}^{Y} 1_{k=y_i} \cdot \text{softmax}\left(W_t^m \cdot \varphi_t^m(\theta^m, x_i^m)\right)_k, \tag{6}$$

where Y is the number of classes. Domain confidence c_i^m is similarly derived from the domain discriminator's output:

$$c_i^m = \sum_{k=1}^{C} 1_{k=d_i} \cdot \text{softmax}\left(D_t^m \cdot \varphi_t^m(\theta^m, x_i^m)\right)_k, \tag{7}$$

where C is the number of domains.

To compare modalities within a mini-batch B_t, we define two discrepancy ratios:

$$\rho_t^m = \frac{\sum_{i \in B_t} q_i^m}{\sum_{i \in B_t} q_i^{\bar{m}}}, \quad \sigma_t^m = \frac{\sum_{i \in B_t} c_i^{\bar{m}}}{\sum_{i \in B_t} c_i^m}, \tag{8}$$

where $\bar{m}$ denotes the other modality. These ratios are defined in complementary directions: high semantic confidence benefits classification (higher numerator), while low domain confidence benefits domain generalization (lower denominator). Hence, $\rho_t^m > 1$ indicates that modality m is stronger in classification, while $\sigma_t^m > 1$ implies that modality m is stronger in domain invariance.

To modulate gradient flow accordingly, we introduce two independent, decoupled modulation coefficients, k_t^m and

p_t^m, for classification and domain discrimination gradients, respectively:

$$k_t^m = \begin{cases} 1 - \tanh(\alpha_k \cdot \rho_t^m) & \text{if } \rho_t^m > 1 \\ 1 & \text{otherwise,} \end{cases} \tag{9}$$

$$p_t^m = \begin{cases} 1 - \tanh(\alpha_p \cdot \sigma_t^m) & \text{if } \sigma_t^m > 1 \\ 1 & \text{otherwise,} \end{cases} \tag{10}$$

where α_k, α_p are hyperparameters controlling suppression strength. The $\tanh$ function ensures coefficients are bounded in $[0, 1]$, maintaining training stability.

This decoupled strategy enables the model to scale down the classification gradient g_c^m using k_t^m and the domain gradient g_d^m using p_t^m, independently. For instance, when modality m is already strong in classification ($\rho_t^m > 1$), k_t^m reduces its classification gradient, allowing weaker modalities to contribute more. A similar principle holds for domain gradients when $\sigma_t^m > 1$.

The final modulated gradients are:

$$\hat{g}_c^m = k_t^m \cdot g_c^m, \quad \hat{g}_d^m = p_t^m \cdot g_d^m. \tag{11}$$

Conflict-Adaptive Gradient Projection

To mitigate inter-task conflicts during optimization, we introduce the Conflict-Adaptive Gradient Projection (CAGP) strategy, which explicitly addresses gradient interference between classification and domain invariant learning objectives. In MMDG, gradient conflicts between these objectives, indicated by $\hat{g}_c^m \cdot \hat{g}_d^m < 0$, are often highly asymmetric across modalities. For example, such conflicts tend to be severe in video ($m = v$) but minimal in audio ($m = a$). Additionally, their relative task strengths vary depending on the modality and over time, making uniform conflict resolution strategies suboptimal.

CAGP tackles these limitations through three strategys. First, it is conflict-aware, projecting only when gradient conflicts arise, which helps maintain task synergy. Second, it is modality specific, projecting independently within each modality to capture distinct dynamics. Third, it is task strength adaptive, favoring updates that support the weaker task to ensure balanced learning and avoid task dominance. Its purpose is to remove conflicting components from the update direction, preserving useful progress while avoiding destructive interference. If $\hat{g}_c^m \cdot \hat{g}_d^m \geq 0$, the gradients are aligned or orthogonal, and no action is needed. To avoid weakening the less optimized task, we prioritize its full update direction. Thus, when a conflict is detected, we resolve it by projecting the stronger task's gradient orthogonal to the weaker one, thereby preserving progress for the weaker objective.

To determine which task is stronger, we use the relative task strength ratio $\Gamma_t^m = \rho_t^m / \sigma_t^m$, where ρ_t^m reflects the classification strength of modality m (via high semantic confidence), and σ_t^m reflects its domain invariance strength (via low domain confidence).

If $\Gamma_t^m > 1$, classification is the relatively stronger task. We project its gradient away from the direction of the domain invariant gradient:

$$\tilde{g}_c^m = \hat{g}_c^m - \frac{\hat{g}_c^m \cdot \hat{g}_d^m}{||\hat{g}_d^m||^2} \hat{g}_d^m. \tag{12}$$

Method	EPIC-Kitchens	HAC
Base	55.06	61.86
Grad-Blending	55.49	62.66
OGM-GE	55.71	62.83
AGM	55.39	62.16
CGGM	55.30	62.80
DRB	54.88	61.92
BALGRAD	55.26	62.36
GMP(Ours)	**57.36**	**64.91**

Table 2: Comparison of our proposed GMP strategy with existing gradient strategies. All methods use concatenation fusion as the base.

This projects g_c^m orthogonally to g_d^m to remove the conflicting component with the domain task. The updated gradient becomes:

$$G_t^m = \tilde{g}_c^m + \hat{g}_d^m. \tag{13}$$

Conversely, if $\Gamma_t^m < 1$, classification is relatively weaker, and we project the domain invariant learning gradient onto a direction orthogonal to the classification task:

$$\tilde{g}_d^m = \hat{g}_d^m - \frac{\hat{g}_d^m \cdot \hat{g}_c^m}{||\hat{g}_c^m||^2} \hat{g}_c^m, \tag{14}$$

and update with:

$$G_t^m = \hat{g}_c^m + \tilde{g}_d^m. \tag{15}$$

This projection strategy ensures that the relatively weaker task's full gradient is preserved, while only the conflicting component of the stronger task is removed.

Experiments

Datasets and Implementation Details. We evaluate on two popular multimodal benchmarks: EPIC-Kitchens (Damen et al. 2018) and HAC (Dong et al. 2023), both offering synchronized video and audio. Following (Dong et al. 2023), we use a domain generalization setup: training on two domains and testing on a third, reporting the average performance across all unseen target domains. Our setup is based on (Dong et al. 2023) and implemented with MMAction2 (Contributors 2020). We use SlowFast for video and ResNet-18 for audio, pretrained on Kinetics-400 (Kay et al. 2017) and VGGSound (Chen et al. 2020), respectively. Models are trained using Adam (learning rate 1e-4) for 20 epochs on a RTX 4090 GPU. The best model is selected based on validation performance, with final results averaged over five runs. Both the domain discriminator and classifier are MLPs with 256 dimensional hidden layers. Hyperparameters are tuned in [0,1] using validation data.

Comparison with Other Gradient Strategies. We compare our method with several state of the art gradient modulation strategies tailored for multimodal learning, including Grad-Blending (Wang, Tran, and Feiszli 2020), OGM-GE (Peng et al. 2022), AGM (Li et al. 2023), CGGM (Guo et al. 2024), DRB (Wei et al. 2024), and BALGRAD (Kwon et al. 2025). To ensure fair comparison, we reimplement

Method	EPIC-Kitchens	HAC
RNA-Net	60.31	62.88
MOOSA	60.29	67.39
SimMMDG	60.81	67.82
CMRF	60.60	67.66
RNA-Net†	60.92	65.21
MOOSA†	61.79	68.75
SimMMDG†	62.03	**69.11**
CMRF†	**62.30**	68.93

Table 3: Integration with existing MMDG methods. Methods marked with † indicate integration of GMP.

Method	EPIC-Kitchens	HAC
Base	55.06	61.86
IGDM-only	55.98	63.05
CAGP-only	55.34	63.41
Full	**57.36**	**64.91**
Unified Modulation	54.97	62.50
w/o k_t^m	55.19	62.33
w/o p_t^m	55.70	62.29
Full IGDM	**55.98**	**63.05**
Fixed Proj-Class	54.63	62.38
Fixed Proj-Domain	53.88	62.03
PCGrad	54.33	62.64
Reverse CAGP	54.00	62.57
Full CAGP	**55.34**	**63.41**

Table 4: Ablation study results demonstrating the effectiveness of individual components and strategies.

all methods using the same architecture and training protocol described above. As shown in Table 2, although existing strategies alleviate gradient imbalance and outperform the naive Base method, they are less effective for MMDG. In contrast, our proposed GMP strategy consistently outperforms all baselines on both EPIC-Kitchens and HAC datasets. These results underscore the need for MMDG specific gradient optimization methods.

Integration with Existing MMDG Methods. To evaluate the compatibility of our optimization strategy with existing MMDG frameworks, we integrate GMP into several representative methods: RNA-NET (Planamente et al. 2024), MOOSA (Dong, Chatzi, and Fink 2024), SimMMDG (Dong et al. 2023), and CMRF (Fan et al. 2024). These baselines primarily target architectural or representational improvements to handle modality and domain gaps. As shown in Table 3, incorporating GMP consistently yields further performance gains across all cases. This demonstrates that our approach functions as a plug-and-play optimization module, enhancing generalization capabilities in unseen domains.

Uni-modal Performance Comparison. As shown in Table 2, our method substantially improves single-modality generalization to unseen domains. The video branch

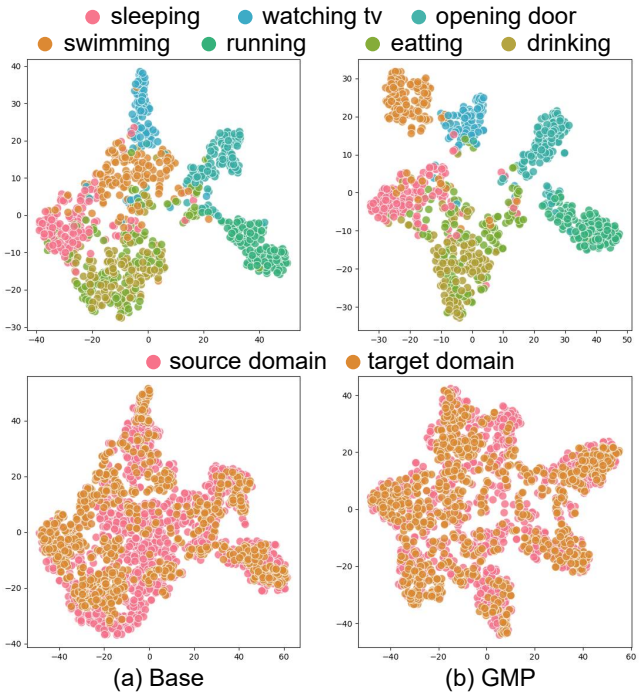

Figure 3: t-SNE plots on the HAC dataset (target domain A) compare the Base and GMP methods in classification and domain invariance.

achieves 52.33% accuracy (a +3.47% gain over the baseline), while the audio branch reaches 35.88% (+1.73%). These results indicate that GMP successfully unlocks the generalization capacity of each modality. Notably, traditional joint training often suppresses unimodal capabilities; for instance, the baseline video branch performs 6.12% worse than its independently trained counterpart. GMP reduces this gap to just 2.65%.

Ablation Study. We perform extensive ablation studies to assess the contribution of each proposed component. Quantitative results are presented in Table 4.

Effectiveness of IGDM and CAGP: Adding IGDM or CAGP individually to the baseline yields noticeable improvements on both EPIC and HAC datasets. When combined, the full model achieves the highest performance, confirming that these components offer complementary benefits.

Decoupled vs. Unified Modulation: Replacing IGDM's decoupled strategy with a unified gradient scaling scheme leads to a performance drop. Further analysis shows that disabling either semantic confidence modulation (k_t^m) or domain confidence modulation (p_t^m) results in significant accuracy loss. These findings emphasize the importance of jointly maintaining both forms of confidence-based modulation for effective modality balancing.

Adaptive Projection vs. Alternatives: Replacing CAGP with fixed projection strategies that target either classification or domain gradients results in inferior performance. Although PCGrad (Yu et al. 2020) partially alleviates conflicts, it still underperforms compared to CAGP. Notably, Reverse CAGP,

Method	EPIC-Kitchens	HAC
Activation-based	52.23	57.49
Concatenation	55.06	61.86
Summation	52.94	58.51
FiLM	53.16	59.35
Activation-based†	55.69	61.33
Concatenation†	**57.36**	**64.91**
Summation†	56.37	63.21
FiLM†	56.19	63.04

Table 5: Combined with GMP, conventional fusion methods consistently gain considerable improvement. † indicates GMP strategy is applied.

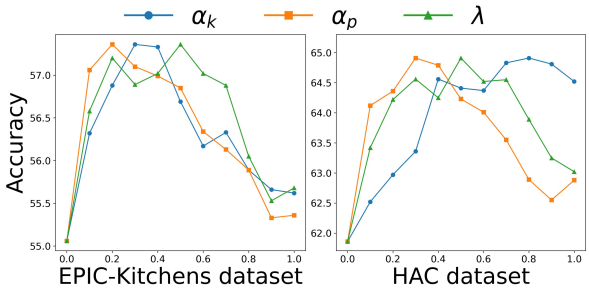

Figure 4: Hyperparameter sensitivity analysis of GMP.

which protects the stronger task rather than the weaker one, also performs worse. These comparisons confirm the design choice of protecting the weaker task through task strength adaptive projection, as this approach more effectively resolves gradient conflicts and improves performance.

Visualization. To further validate our approach, we present a t-SNE visualization on the HAC dataset (target domain A) in Fig. 3. As observed, the baseline method suffers from overlapping class clusters and a distinct distribution gap between domains. Conversely, our GMP method produces well-separated decision boundaries with distinct class structures. Furthermore, GMP demonstrates superior domain invariance, evidenced by the significant alignment between source and target features. This qualitative analysis confirms that our dual-objective strategy effectively learns representations that are both class-discriminative and domain-invariant, ensuring robust generalization.

Integration with Conventional Fusion Methods. To further demonstrate the versatility of our approach, we integrate GMP with various conventional fusion strategies, including summation, concatenation, activation-based fusion, and the adaptive FiLM mechanism (Perez et al. 2018). As shown in Table 5, our method consistently improves performance across all fusion types. These results demonstrate that GMP is agnostic to fusion paradigms and complements a wide range of multimodal fusion schemes, enhancing generalization without reliance on any specific fusion design.

Hyperparameter Sensitivity. We conduct controlled experiments to evaluate the sensitivity of GMP to its hyperparameters α_k, α_p, and λ. The parameters α_k and α_p control the

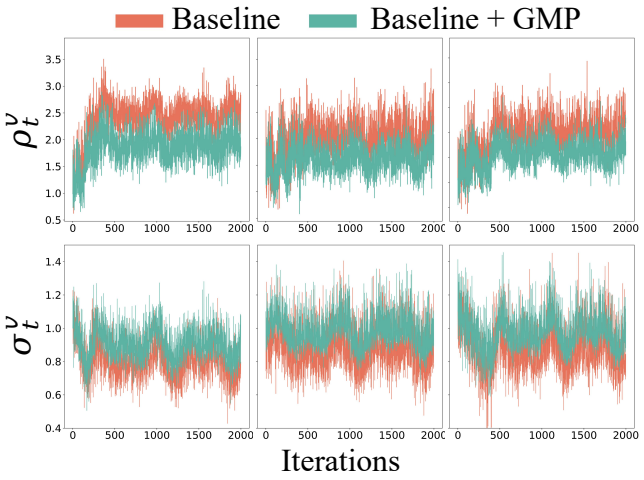

Figure 5: Discrepancy ratio trajectories (ρ_t^m, σ_t^m) on EPIC-Kitchens. GMP mitigates cross-modal imbalance by driving ratios toward 1, achieving balanced semantic and domain optimization.

suppression strength of dominant modalities in the classification and domain branches, respectively, while λ balances classification accuracy with domain invariance. Setting α_k and α_p to 0.0 disables suppression, whereas higher values apply stronger attenuation. As shown in Fig. 4, most settings consistently improve performance over the baseline, indicating that calibrated suppression reduces modality dominance and improves generalization without diminishing the influence of strong modalities.

Discrepancy Ratio Analysis. To quantify the balance between modalities during training, we define two discrepancy ratios: the semantic discrepancy ratio (ρ_t^m) and the domain discrepancy ratio (σ_t^m), as described in Eq. (8). These ratios measure the relative semantic discriminativeness and domain invariance of modality m with respect to its counterpart $\bar{m}$ at training step t. Ideally, both values converge to 1, reflecting balanced contributions. Significant deviations indicate optimization asymmetry, where one modality dominates, leading to underutilization of the other. As illustrated in Fig. 5, training without our method results in persistent imbalance. In contrast, our GMP reduces this discrepancy by guiding both ρ_t^m and σ_t^m closer to 1, thereby promoting balanced inter-modality learning.

Conclusion

We propose Gradient Modulation Projection (GMP), a unified optimization strategy for Multimodal Domain Generalization (MMDG). GMP addresses gradient imbalances and task conflicts in multimodal learning, enabling models to better generalize to unseen domains. Experiments on multiple datasets demonstrate state-of-the-art generalization performance. GMP is architecture-agnostic, consistently enhances existing MMDG methods, and offers the first optimization-centric solution for multimodal domain shifts.

Acknowledgments

This work was supported by the National Key R&D Program of China (2024YFB3311600) and the Natural Science Foundation of Henan (252300423936).

References

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)*, 721–725. IEEE.

Contributors, M. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, 720–736.

Dong, H.; Chatzi, E.; and Fink, O. 2024. Towards Multimodal Open-Set Domain Generalization and Adaptation through Self-supervision. In *European Conference on Computer Vision*.

Dong, H.; Liu, M.; Zhou, K.; Chatzi, E.; Kannala, J.; Stachniss, C.; and Fink, O. 2025a. Advances in Multimodal Adaptation and Generalization: From Traditional Approaches to Foundation Models. *arXiv preprint arXiv:2501.18592*.

Dong, H.; Nejjar, I.; Sun, H.; Chatzi, E.; and Fink, O. 2023. SimMMDG: A simple and effective framework for multimodal domain generalization. *Advances in Neural Information Processing Systems*, 36: 78674–78695.

Dong, H.; Sheng, L.; Liang, J.; He, R.; Chatzi, E.; and Fink, O. 2025b. Adapting Vision-Language Models Without Labels: A Comprehensive Survey. *arXiv preprint arXiv:2508.05547*.

Fan, Y.; Xu, W.; Wang, H.; and Guo, S. 2024. Cross-modal representation flattening for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 37: 66773–66795.

Galappaththige, C. J.; Baliah, S.; Gunawardhana, M.; and Khan, M. H. 2024. Towards Generalizing to Unseen Domains with Few Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23691–23700.

Galappaththige, C. J.; Kuruppu, G.; and Khan, M. H. 2024. Generalizing to unseen domains in diabetic retinopathy classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7685–7695.

Guo, Z.; Jin, T.; Chen, J.; and Zhao, Z. 2024. Classifier-guided gradient modulation for enhanced multimodal learning. *Advances in Neural Information Processing Systems*, 37: 133328–133344.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Khan, A.; Shaaban, M. A.; and Khan, M. H. 2024. Improving pseudo-labelling and enhancing robustness for semi-supervised domain generalization. *arXiv preprint arXiv:2401.13965*.

Kwon, J.; Kim, M.; Lee, E.; Choi, J.; and Kim, Y. 2025. See-Saw Modality Balance: See Gradient, and Sew Impaired Vision-Language Balance to Mitigate Dominant Modality Bias. In *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)*, 4364–4378.

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*, 22214–22224.

Li, H.; Liu, S.; Wang, H.; Jiang, X.; Jiu, M.; Chen, L.; Lu, Y.; Li, S.; and Xu, M. 2025a. RRGMambaFormer: A hybrid Transformer-Mamba architecture for radiology report generation. *Expert Systems with Applications*, 279: 127419.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Li, H.; Wan, H.; Zhang, L.; Jiu, M.; Li, S.; Xu, M.; and Khan, M. H. 2025b. Towards Robust Multimodal Domain Generalization via Modality-Domain Joint Adversarial Training. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 180–188.

Li, H.; Wang, H.; Sun, X.; He, H.; and Feng, J. 2025c. Context-enhanced framework for medical image report generation using multimodal contexts. *Knowledge-Based Systems*, 310: 112913.

Li, Y.; Cao, Y.; He, H.; Cheng, Q.; Fu, X.; Xiao, X.; Wang, T.; and Tang, R. 2025d. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*.

Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; and Tang, R. 2025e. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 736–763.

Ma, F.; Liu, L.; and Cheng, H. V. 2024. TIMA: Text-Image Mutual Awareness for Balancing Zero-Shot Adversarial Robustness and Generalization Ability. *arXiv preprint arXiv:2405.17678*.

Munir, M. A.; Khan, M. H.; Sarfraz, M. S.; and Ali, M. 2023. Domain adaptive object detection via balancing between self-training and adversarial learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14353–14365.

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*, 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 AAAI conference on artificial intelligence*, volume 32.

Planamente, M.; Plizzari, C.; Alberti, E.; and Caputo, B. 2022. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1807–1818.

Planamente, M.; Plizzari, C.; Peirone, S. A.; Caputo, B.; and Bottino, A. 2024. Relative norm alignment for tackling domain shift in deep multi-modal classification. *International Journal of Computer Vision*, 132(7): 2618–2638.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Philip, S. Y. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8): 8052–8072.

Wang, W.; Tran, D.; and Feiszli, M. 2020. What Makes Training Multi-Modal Classification Networks Hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, Y.; Li, S.; Feng, R.; and Hu, D. 2024. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, 71–86. Springer.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4396–4415.