Reconcile Gradient Modulation for Harmony Multimodal Learning

Xiyuan Gao[1,2,3], **Bing Cao**[1,4*], **Baoquan Gong**[1], **Pengfei Zhu**[1,2,3*]

[1]School of Artificial Intelligence, Tianjin University, Tianjin, China
[2]Low-Altitude Intelligence Lab, Xiong'an National Innovation Center, Xiong'an, China
[3]Xiong'an Guochuang Lantian Technology Co., Ltd., Xiong'an, China
[4]Haihe Lab of ITAI, Tianjin, China
{gaoxiyuan, caobing, gongbaoquan, zhupengfei}@tju.edu.cn

Abstract

Multimodal learning frequently faces two coupled challenges: *modality imbalance*, where dominant modalities suppress others during training, and *modality conflict*, where opposing gradient directions hinder optimization. Existing methods typically address these issues in isolation, yet they are intrinsically correlated and most fundamentally reflected in the gradient space—severe imbalance may obscure conflicts, while suppressing conflict may homogenize features and worsen imbalance, affecting fusion performance. To jointly address this coupled challenge, we propose Reconcile Gradient Modulation (RGM), a unified framework that adaptively adjusts gradient magnitude and direction for harmony multimodal learning. The core of RGM is Syn-Orth Grad, which minimizes Dirichlet energy to perform minimal-gradient surgery. It enhances cooperation synergy when modalities are aligned and enforces orthogonality to preserve uniqueness in conflict situations, thus promoting stable and balanced learning. To guide this modulation, we propose Cumulative Gradient Energy (CGE) as a convergence-guaranteed measure of modality-wise progress, and construct a Balance-nonConflict Plane (BCP) for real-time diagnosis and control of training dynamics. Experiments on diverse benchmarks validate our effectiveness and generalizability, consistently outperforming counterparts that are designed to handle multimodal imbalance or conflict independently.

Introduction

Multimodal learning has emerged as a pivotal paradigm for integrating heterogeneous information sources across a wide array of real-world applications (Gallo, Calefati, and Nawaz 2017; Hong et al. 2025; Tzirakis et al. 2021; Zhang et al. 2025). By combining complementary modalities, multimodal models typically surpass unimodal baselines (Huang et al. 2021). However, in practice, the expected synergy often fails to materialize due to two persistent challenges: modality imbalance and modality conflict. The former refers to the phenomenon where certain modalities dominate the training process, leading to insufficient representation learning from weaker modalities (Wu et al. 2022). The latter denotes the representational inconsistency or opposition between modalities, often manifested as low feature similarity or even antagonistic gradient directions during optimization (Sener and Koltun 2018).

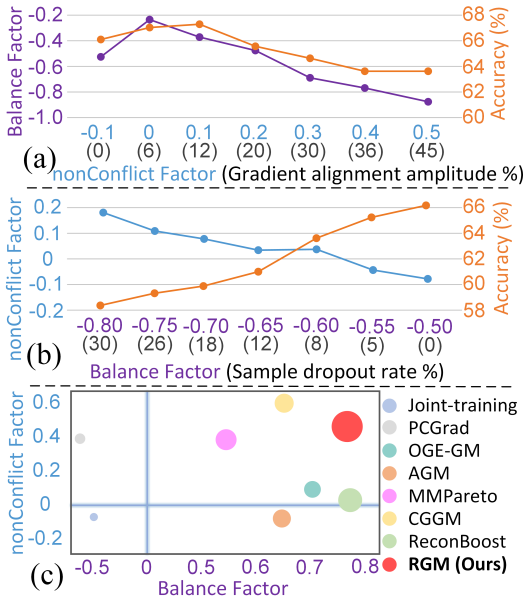

Figure 1: **(a)&(b)** Interplay between modality imbalance and gradient conflict. (a) Conflict-only modulation amplifies imbalance; (b) Balance-only modulation led to more conflicts. **(c)** Balance-nonConflict plane quality comparison on the KS dataset. A larger spot size reflects a higher accuracy.

A variety of techniques have been proposed to address these issues (Imfeld et al. 2023; Li et al. 2023b). For modality imbalance, gradient modulation approaches such as OGM-GE (Peng et al. 2022) and AGM (Li et al. 2023a) aim to rescale the magnitude of gradients from each modality in real time, promoting a more balanced learning pace. Alternatively, unimodal assistance frameworks (Wei et al. 2024b) integrate auxiliary unimodal objectives to preserve modality-specific information, indirectly enhancing underutilized modalities. However, these additional objectives often introduce extra gradient conflicts between unimodal and multimodal losses, complicating optimization. To mitigate

*Corresponding author

such conflicts, MMPareto (Wei and Hu 2024) attempt to reconcile gradient directions across modalities via Pareto optimization. Despite their merits, a core limitation remains: most existing methods treat imbalance and conflict as isolated problems, failing to recognize their interdependent nature. As shown in Fig. 1(a), forcibly aligning the *Audio* (dominant) and *Video* (weaker) modalities to increase similarity and reduce conflict leads to feature homogenization, which undermines the distinctiveness of each modality and consequently exacerbates imbalance. While, in Fig. 1(b), we simulate varying degrees of modality imbalance by changing the dropout rate of the *Video* modality. Rising balance factor cause more conflicts, and severe imbalance suppresses weaker modality, hiding gradient conflicts. Although some methods like CGGM (Guo et al. 2024) attempt to jointly modulate both gradient magnitude and direction, their conflict resolution strategies often rely on naively increasing feature similarity across modalities, which risks erasing modality-specific information vital for robust fusion.

To overcome these limitations, we propose Reconcile Gradient Modulation (RGM), a unified framework that adaptively modulates both gradient magnitude and direction based on cross-modal agreement, without relying on additional unimodal losses. This enables dynamic cooperation or separation, achieving a fine-grained balance between fusion and specialization. Crucially, to support this dynamic modulation, we introduce a novel diagnostic tool, the Balance-nonConflict Plane (BCP), which provides a continuous, interpretable view of the current multimodal optimization state. The coordinates on the BCP reflect intermodal learning balance and the degree of representational conflict across modalities. Building upon this geometric intuition, we develop a dynamic yet effective gradient modulation called SynOrth Grad, which uses the coordinates on the BCP to synchronize the resolution of modality imbalance and conflict in a seamless manner. Without the need for hand-crafted weighting schemes or extra supervisory signals, SynOrth Grad can adaptively strengthen the degree of gradient alignment or orthogonality, and is independent of both the model architecture and feature fusion strategy. The main highlights of our study are as follows:

- We propose Reconcile Gradient Modulation (RGM), the first framework to simultaneously and adaptively regulate modality imbalance and conflict from a unified perspective, thereby effectively promoting balanced and coordinated multimodal optimization.
- To implement RGM, we develop the SynOrth Grad formulation, which minimizes Dirichlet energy to perform synergistic enhancement and orthogonal rotation of gradients with minimal distortion. This leads to effective cooperation between aligned modalities and disentanglement between conflicting ones.
- We further introduce Cumulative Gradient Energy (CGE) as a convergence-guaranteed indicator of modality-wise learning progress, and construct a Balance-nonConflict Plane for real-time training diagnosis, providing interpretability and consistently boosting performance across multiple multimodal benchmarks.

Related Work

Multimodal learning often faces two persistent challenges that hinder the performance of models: modality imbalance and modality conflict. Modality imbalance arises when dominant modalities overpower weaker ones during training, resulting in underutilization of certain modality-specific representations. Modality conflict occurs when different modalities produce incompatible features or opposing gradients, thereby disrupting effective fusion. Both issues are prevalent yet often treated independently in the literature.

Imbalance in multimodal learning

Previous studies have proposed several techniques. Gradient-based modulation methods (Jiang, Chi, and Yang 2025; Huang et al. 2021) such as OGM-GE (Peng et al. 2022) dynamically rescale gradients to equalize learning progress among modalities. Others, like PMR (Fan et al. 2023) and AGM (Li et al. 2023a), design modality-specific evaluation strategies, e.g., clustering tightness (Wei et al. 2024b) or Shapley-based (Hu, Li, and Zhou 2022; Wei et al. 2024a) contribution to guide balanced updates. Recent efforts also explore theoretical modeling of modality competition and saturation (Huang et al. 2025). Nevertheless, these approaches commonly assess imbalance based on task-specific metrics such as accuracy, which reflect final performance but not learning dynamics. Such metrics cannot reveal whether a modality is truly under-optimized or simply less informative. In contrast, we propose a Cumulative Gradient Energy metric that captures each modality's learning state over time. This measure is architecture-agnostic, correlates with loss descent, and is theoretically shown to align with convergence guarantees, which offer a more robust and general perspective on modality imbalance.

Conflict in multimodal learning

Recent work has shifted from passive observation to active gradient regulation. MMPareto (Wei and Hu 2024) formulates multi-objective optimization to resolve conflicts between multimodal and unimodal losses; CGGM (Guo et al. 2024) leverages classifier heads to modulate gradient directions across modalities. ReconBoost (Hua et al. 2024) introduces stage-wise alternating optimization to reduce mutual interference between competing modalities. MMCosine (Xu et al. 2023) dynamically adjusts gradient similarity to ensure cooperative updates, while BALGRAD (Kwon et al. 2025) introduces a soft-switch mechanism to promote underused modalities when conflicts emerge. MLA (Zhang et al. 2024) and Remixing (Ma, Chen, and Deng 2025) mainly alternate unimodal learning, avoiding cross-modal conflict.These methods have demonstrated success but often rely on auxiliary losses, branches, or sample-specific fusion, which increase model complexity and may amplify new forms of conflict. In contrast, our method operates directly on the shared gradient space, without introducing extra parameters or losses. It achieves minimal-disruption dynamic modulation, aligning gradients based on observed conflict levels and naturally balancing specialization and cooperation. Furthermore, our method is compatible with arbitrary architecture and feature fusion strategies, making it widely applicable.

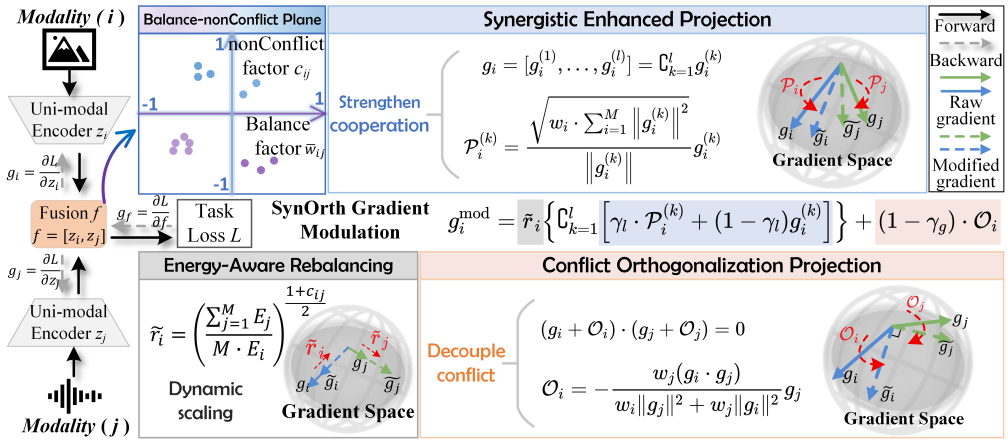

Figure 2: Overview of Reconcile Gradient Modulation. A standard forward–backward pass yields raw gradients for each modality. We compute Cumulative Gradient Energy ($E_{i,j}$) and cosine similarity to get Balance factor ($\bar{w}_{ij}$) and nonConflict factor(c_{ij}), respectively, mapping the current learning state to the Balance-nonConflict Plane. According to the plane coordinates, SynOrth Gradient Modulation performs synergistic enhancement ($\mathcal{P}_{i,j}$) and orthogonal rotation ($\mathcal{O}_{i,j}$) under the control of threshold γ, then applies dynamic scaling of gradient magnitudes ($\tilde{r}_{i,j}$) to balance modality contributions, producing minimally modified gradients that are fed back to the optimizer. This unified procedure simultaneously alleviates modality imbalance and conflict.

Methods

Preliminary

We consider a standard *multi-modal learning* setup with M modalities $\{x_1, \ldots, x_M\}$, each encoded by a modality-specific encoder $E_k(\cdot)$ into features $z_k = E_k(x_k) \in \mathbb{R}^{d_k}$. These are combined via a fusion function $F(\cdot)$ to produce a joint representation:$f = F(z_1, \ldots, z_M) \in \mathbb{R}^d$, which is then passed through a task head $h(\cdot)$ to generate prediction $\hat{y} = h(f)$. Gradient notation. We optimize the model using a loss function $\mathcal{L}(\hat{y}, y)$ with parameters θ in the fusion and prediction head. Let $g = \nabla_f \mathcal{L} \in \mathbb{R}^d$ be the gradient backpropagated to the fused feature f. To analyze modality contributions, we compute the gradient with respect to each modality feature z_k using the chain rule:

$$g_k = \nabla_{z_k} \mathcal{L} = \frac{\partial \mathcal{L}}{\partial f} \cdot \frac{\partial f}{\partial z_k} = g \cdot J_k, \tag{1}$$

where $J_k = \frac{\partial f}{\partial z_k} \in \mathbb{R}^{d \times d_k}$ is the Jacobian of the fusion function. For concatenation-based fusion, g_k corresponds to a slice of g, i.e., $g_k = g[s_k : s_k + d_k]$ with $s_k = \sum_{i<k} d_i$.

Cumulative Gradient Energy

In multimodal learning, dominant modalities often dominate parameter updates, while weaker ones remain under-optimized. To fundamentally quantify each modality's contribution, previous works (Achille and Soatto 2018; Huang et al. 2025) use Fisher Information Matrix (FIM) (Fisher 1925), which models local curvature via:

$$F_k = \mathbb{E}_{x_k} \left[\nabla_\theta \mathcal{L}(x_k, y) \, \nabla_\theta \mathcal{L}(x_k, y)^\top \right]. \tag{2}$$

Despite its theoretical appeal, FIM is second-order, static, and insensitive to long-term influence.

Inspired by Synaptic Intelligence (Zenke, Poole, and Ganguli 2017), which accumulates the alignment between gradient and actual parameter change over time. We propose the *Cumulative Gradient Energy* (CGE) as a dynamic alternative. Let $g_k^{(t)}$ denote the gradient from modality k at iteration t, and η_t is the effective learning rate (reflecting learning rate schedule and optimizer dynamics). Based on the standard SGD-style update rule: $\theta^{(t+1)} = \theta^{(t)} - \eta_t \cdot g_k^{(t)}$, the total parameter shift due to modality k up to iteration T is: $\theta^{(T)} - \theta^{(0)} = -\sum_{t=1}^{T} \eta_t \cdot g_k^{(t)}$, CGE of modality k is:

$$E_k^{(T)} = \sum_{t=1}^{T} \left\| \eta_t \cdot g_k^{(t)} \right\|^2, \tag{3}$$

this quantity reflects the total *work* done by modality k, an intuitive, energy-based measure of its role in optimization.

Theorem 1. CGE is not merely descriptive; under standard assumptions (complete proof process is provided in the *Appendix*), it aligns with the convergence of training:

$$E_k^{(T)} \lesssim 2\eta \cdot \left[\mathcal{L}(\theta^{(0)}) - \mathcal{L}(\theta^{(T)}) \right], \tag{4}$$

where η is a constant learning rate during training. This ensures CGE faithfully reflects actual optimization progress without drift or saturation. In summary, CGE provides an optimizer-aware, convergence-consistent, and physically meaningful estimate of modality's accumulated contribution, forming the basis of our gradient modulation strategy.

Balance–nonConflict Plane

To characterize the interplay between *modality imbalance* and *modality conflict*, we construct a compact 2D representation: the *Balance-nonConflict Plane* (BCP). Each modality pair (i, j) is mapped to a point $(\bar{w}_{ij}, c_{ij}) \in [-1, 1]^2$, where: **Balance axis.** We use CGE to compute modality contributions: $w_i = \frac{E_i}{E_i + E_j}$, and define the balance coefficient:

$$\bar{w}_{ij} = 8 w_i w_j - 1 \in [-1, 1], \tag{5}$$

where $\bar{w}_{ij} = 1$ indicates perfect balance ($w_i = w_j = 0.5$), and $\bar{w}_{ij} = -1$ reflects extreme imbalance.

nonConflict axis. We compute the cosine similarity between modality's gradients:

$$c_{ij} = \frac{g_i^\top g_j}{\|g_i\| \cdot \|g_j\|} \in [-1, 1], \tag{6}$$

where $c_{ij} = 1$ indicates the most synergistic and $c_{ij} = -1$ implies the most severe conflict. This 2D plane offers an interpretable view of modality interactions and guides the adaptive gradient modulation introduced in the next section.

Reconcile Gradient Modulation

Energy-Aware Rebalancing. To mitigate modality imbalance, we compute a CGE-based scaling factor:

$$r_i = \frac{\sum_{j=1}^{M} E_j}{M \cdot E_i}, \tag{7}$$

This formulation ensures that under-contributing modalities (with lower E_i) receive stronger gradient amplification, while dominant ones are gently suppressed, encouraging more balanced parameter updates. To avoid destabilization in conflicting cases, we make the scaling conflict-aware by modulating r_i with nonConflict factor c_{ij}:

$$\tilde{r}_i = r_i^{\frac{1+c_{ij}}{2}}, \tag{8}$$

This design makes the scaling more responsive when gradients are aligned ($c_{ij} \to 1$), preventing the exacerbation of imbalances caused by concealed conflicts, while reducing modulation strength when gradients are contradictory ($c_{ij} \to -1$). This prevents conflict from amplifying imbalance and stabilizes training in highly imbalance condition.

Synergistic Enhanced Projection. When the gradients of two modalities exhibit partial alignment, we encourage further cooperation by aligning their directions locally.

Theorem 2. *Encouraging proportional scaling within corresponding local blocks of two vectors (i.e., dividing each vector into k parts) can provably improve their overall cosine similarity.* (The proof process is provided in the *Appendix.*)

Based **Theorem 2**, we divide each modality gradient into l blocks: $g_i = \mathbb{C}_{k=1}^{l} g_i^{(k)} = [g_i^{(1)}, \ldots, g_i^{(l)}]$. For each block k, we aim to make $\|g_i^{(k)}\|/\|g_j^{(k)}\| \approx \rho$,where ρ is a constant, to achieve collinearity. We compute scaling coefficients $\alpha_i^{(k)}$ by minimizing the general Dirichlet energy, a classical smoothness functional that penalizes variations between paired gradient blocks:

$$\min_{\alpha_i^{(k)}} \mathcal{D}_{\text{dir}} = \sum_{k=1}^{l} \left\| \mathcal{P}_i^{(k)} - g_i^{(k)} \right\|^2, \quad \mathcal{P}_i^{(k)} = \alpha_i^{(k)} g_i^{(k)} \tag{9}$$

$$\text{s.t.} \quad \begin{cases} \|\mathcal{P}_i^{(k)}\|/\|\mathcal{P}_j^{(k)}\| = \rho, \\ \sum_{i=1}^{M} \|\mathcal{P}_i^{(k)}\| = \sum_{i=1}^{M} \|g_i^{(k)}\|. \end{cases} \tag{10}$$

The closed-form solution is:

$$\alpha_i^{(k)} = \frac{\sqrt{w_i \cdot \sum_{i=1}^{M} \|g_i^{(k)}\|^2}}{\|g_i^{(k)}\|} \tag{11}$$

This formulation yields the least-distorting alignment across modalities, enhancing their synergy in a block-wise manner without aggressive correction. In summary, this synergy-promoting adjustment helps unify gradient directions at a fine-grained level, improving modality cooperation while minimizing structural interference. The derivation of this result is provided in the *Appendix.*

Conflict Orthogonalization Projection. To alleviate destructive gradient conflict ($c_{ij} < 0$), we seek to rotate the gradients of the conflicting modalities to become orthogonal, while minimizing deviation from their original directions. Let $\tilde{g}_i = g_i + \mathcal{O}_i$, we aim to achieve $\tilde{g}_i^\top \tilde{g}_j \approx 0$, where $\mathcal{O}_i$ denotes the adjustment. We minimize the Dirichlet energy $\mathcal{D}_{\text{dir}} = \sum_{i=1}^{M} \|\mathcal{O}_i\|^2$ under an orthogonality constraint:

$$\min_{\mathcal{O}_i, \mathcal{O}_j} \left(w_i \|\mathcal{O}_i\|^2 + w_j \|\mathcal{O}_j\|^2 \right) \quad \text{s.t.} \quad \tilde{g}_i^\top \tilde{g}_j = 0. \tag{12}$$

The contribution weights w_i and w_j guide the responsibility of each modality in the adjustment: dominant modalities are rotated more, while weaker ones retain stability. The closed-form solution is:

$$\mathcal{O}_i = -\frac{w_j (g_i^\top g_j)}{w_i \|g_j\|^2 + w_j \|g_i\|^2} g_j. \tag{13}$$

This formulation reduces gradient interference by promoting directional separation, while preserving modality intent in proportion to their optimization impact. Full derivation is given in the *Appendix.*

Synergistic-Orthogonal Gradient Modulation. As shown in Fig. 2, while energy-aware rebalancing and two directional projections (synergistic & orthogonal) each target different aspects of multimodal conflict, they are not mutually exclusive. In fact, they operate at different spatial resolutions: synergy projection works at the block level to encourage local co-alignment, whereas orthogonalization targets global direction disentanglement when severe conflict. To fully leverage this, we propose a unified Synergistic Orthogonal Gradient modulation (SynOrth Grad).

To enable smooth transitions between cooperation and disentanglement, we introduce soft gates based on the coordinate values on the BCP:

$$\begin{cases} \gamma_g = \sigma\left(q \cdot c_{ij}\right), \\ \gamma_l = \sigma\left(q \cdot c_{ij}^{(k)}\right), \end{cases} \quad \sigma(x) = \frac{1}{1+e^{-x}}, q = 4\bar{w}_{ij}, \tag{14}$$

where c_{ij} is the global cosine similarity, $c_{ij}^{(k)}$ is the block-wise similarity, and $\bar{w}_{ij}$ is the balance factor. The slope q increases with modality balance, sharpening gate transitions. The final modulated gradient for modality i is:

$$g_i^{\text{mod}} = \tilde{r}_i \left\{ \mathbb{C}_{k=1}^{l} \left[\gamma_l \cdot \mathcal{P}_i^{(k)} + (1-\gamma_l) g_i^{(k)} \right] \right\} + (1-\gamma_g) \cdot \mathcal{O}_i. \tag{15}$$

Here, the scaling factor $\tilde{r}_i$ derived from CGE is applied only to the synergy projection and not to the orthogonal residual. This ensures that cooperation is strengthened where possible without forcing uncooperative gradients into unnatural alignment. This unified scheme, SynOrth Grad integrates directional coordination with energetic rebalancing in a fully differentiable and data-driven manner, without relying on hard thresholds or handcrafted heuristics.

Methods	CNN-based Model						Transformer-based Model					
	Kinetics Sounds		CREMA-D		UCF-51		Kinetics Sounds		CREMA-D		UCF-51	
	Acc (%)	$\bar{w}_{ij}$	Acc(%)	$\bar{w}_{ij}$	Acc(%)	$\bar{w}_{ij}$	Acc(%)	$\bar{w}_{ij}$	Acc(%)	$\bar{w}_{ij}$	Acc(%)	$\bar{w}_{ij}$
Joint training	65.74	-0.56	67.47	-0.60	68.23	-0.75	51.28	-0.48	55.74	-0.56	71.62	-0.80
G-Bleeding (CVPR 2020)	68.33	0.61	69.11	0.39	69.14	0.24	55.14	0.58	58.47	0.39	75.16	0.15
Greedy (ICML 2022)	65.72	-	68.37	-	71.53	-	52.16	-	53.16	-	73.53	-
OGM-GE (CVPR 2022)	67.15	0.71	70.70	0.54	71.66	0.40	54.21	0.76	56.21	0.55	77.10	0.28
AL (ACM MM 2023)	66.35	0.26	67.81	0.19	69.51	0.12	49.05	0.33	50.62	0.20	72.36	-0.07
IMCL (ACM MM 2023)	65.92	-	68.17	-	67.46	-	51.95	-	52.61	-	70.21	-
AGM (ICCV 2023)	66.50	0.65	70.16	0.57	72.38	0.40	53.23	0.71	56.14	0.59	76.64	0.42
PMR (CVPR 2023)	65.62	0.67	68.55	<u>0.62</u>	74.80	0.41	53.86	0.66	57.55	0.58	76.05	0.38
Diag&Re (ECCV 2024)	69.10	0.73	**75.13**	<u>0.62</u>	74.55	<u>0.55</u>	58.73	0.75	62.12	0.66	77.08	<u>0.54</u>
LFM (NeurIPS 2024)	69.05	-	72.15	-	75.62	-	57.51	-	61.14	-	76.19	-
ARM (AAAI 2025)	69.65	<u>0.76</u>	73.16	0.59	75.60	0.51	58.64	**0.82**	61.06	<u>0.67</u>	76.25	0.49
Remixing (ICML 2025)	71.68	0.70	72.72	0.61	75.49	0.52	58.49	0.76	<u>62.13</u>	**0.68**	<u>77.81</u>	0.50
InfoReg (CVPR 2025)	<u>72.03</u>	0.64	71.90	0.53	<u>76.01</u>	0.50	<u>59.04</u>	0.69	61.93	0.61	77.39	0.47
RGM (Ours)	**72.56**	**0.77**	<u>74.91</u>	**0.65**	**76.30**	**0.57**	**60.19**	<u>0.78</u>	**62.47**	<u>0.67</u>	**78.69**	**0.55**

Table 1: Comparison with imbalanced methods on three datasets under CNN-based and Transformer-based backbones. For fair comparison, we reproduce all baselines under the same training protocol and report both the *top-1 accuracy (Acc)* and our proposed *Balance Factor* $\bar{w}_{ij}$, which quantifies the relative contribution strength between modalities during training (higher is better). Bold and underline represent the **best** and <u>second best</u> respectively.

Experiments

Dataset and Experiment Settings

Kinetics Sounds (Arandjelovic and Zisserman 2017) is selected 31 categories from the Kinetics-400 dataset (Kay et al. 2017), contains paired video and audio recordings of human-object interactions. The dataset contains approximately 20k video slices, with a ratio of approximately 8:1:1 between the training set, validation set, and test set. **CREMA-D** (Cao et al. 2014) provides 7,442 clips of 91 actors expressing six emotions across audio and visual modalities. The entire dataset is randomly divided into a train and test set of 6,698 and 744 clips, respectively. **UCF-51** is a challenging subset of UCF-101 (Soomro, Zamir, and Shah 2012), includes 6,845 video clips covering 51 action categories selected for higher inter-class similarity, offering both RGB and optical flow (OF) sequences. We adopted the official split 5,466 clips for training and 1,379 for testing.

Experiment Settings. We employ CNN-based (ResNet-18) (He et al. 2016) and Transformer-based (ViT-Base) (Dosovitskiy et al. 2020) as the backbone by default, unless otherwise specified. Followed the InfoReg (Huang et al. 2025) setup, unimodal features are integrated with static feature-level fusion method. For late-fusion baselines in the comparative experiments, we follow the original settings proposed in their respective papers, where unimodal gradients are directly computed from unimodal loss functions. For the Kinetics Sounds and CREMA-D datasets, all models are trained from scratch to ensure fair evaluation. In contrast, models on UCF-51 are initialized with ImageNet-pretrained weights, following common practice for video classification benchmarks (Wang et al. 2018; Zhang, Hao, and Ngo 2021). Training settings closely follow prior works to ensure comparability, all training is conducted using stochastic gradient descent (SGD) with a momentum of 0.9, learning rate of $1e-3$, weight decay of $1e-4$, and a batchsize of 64. Extended experiments are provided in the *Appendix*.

Comparison with related imbalanced methods

Table 1 summarizes the performance of our proposed method RGM against 13 representative baselines under 3 datasets (KS, CREMA-D, UCF-51) and two backbones (CNN-based and Transformer-based). The compared methods fall into 3 categories: (1) naive fusion approaches (Joint training) utilizes concatenation fusion with a single multimodal cross-entropy loss function; (2) loss/gradient weighting techniques (G-Bleeding (Huang et al. 2021), AGM (Li et al. 2023a), ARM (Gao et al. 2025), Diag&Re (Wei et al. 2024b), PMR (Fan et al. 2023)), which balance modalities via manually or adaptively tuned importance; and (3) information modulation methods (Greddy (Wu et al. 2022), OGM-GE (Peng et al. 2022), AL (Shen et al. 2023), IMCL (Zhou et al. 2023), LFM (Yang et al. 2024), Remixing (Ma, Chen, and Deng 2025), InfoReg (Huang et al. 2025)) focusing on representational alignment or regularization.

RGM consistently achieves the best performance across all settings. In terms of accuracy, it improves over the strongest baselines by +0.53% on KS and +0.29% on UCF-51 under CNN-based backbones. In addition, RGM achieves the highest balance factor $\bar{w}_{ij}$, with gains up to +0.06–0.10 over the best prior methods. These improvements reflect not only better task performance but also more equitable gradient contribution across modalities—mitigating modality collapse and improving generalization.

Interestingly, we observe that Transformer backbones without pretraining tend to exhibit better balance (higher w_{ij}) than CNNs, possibly due to faster knowledge acquisition. However, once pretrained weights are loaded—as in UCF-51—the strong prior in visual modality may bias learning, leading to decreased $\bar{w}_{ij}$ and increased modality dominance. This highlights the need for RGM, which promotes cooperation when beneficial and decoupling when necessary. Moreover, by adaptively scaling gradients with training progress, RGM mitigates modality collapse and ensures

Methods	CNN			Transformer		
	KS	CD	UCF	KS	CD	UCF
Joint training	65.74	67.47	68.23	51.28	55.74	71.62
GradNorm (ICML 2018)	65.84	68.96	70.18	56.02	57.37	73.11
PCGrad (NeurIPS 2020)	69.11	70.04	71.22	56.17	58.63	73.31
MMCos (ICASSP 2023)	65.48	66.40	70.06	55.31	57.29	71.01
GMD (AAAI 2024)	67.26	71.49	73.12	57.85	60.26	75.97
ReconB (ICML 2024)	71.34	74.22	75.55	58.14	61.73	77.10
MMPareto (ICML 2024)	70.13	**75.13**	75.03	58.36	60.27	76.69
CGGM (NeurIPS 2024)	70.92	72.13	74.89	59.82	61.65	77.62
BALRAD (NAACL 2025)	69.49	73.45	75.63	58.63	61.36	78.13
RGM (Ours)	**72.56**	74.91	**76.30**	**60.19**	**62.47**	**78.69**

Table 2: Comparison with conflict-aware modulation methods on Kinetics Sounds **(KS)**, CREMA-D **(CD)**, and UCF-51 **(UCF)** datasets.

EAR	Syn	Orth	KS	CREMA-D	UCF-51
			65.74	67.47	68.23
✓			68.13	70.27	71.28
	✓		68.01	69.85	69.53
		✓	67.52	69.26	72.14
✓	✓		71.69	73.51	73.09
✓		✓	70.54	71.64	74.59
	✓	✓	70.97	72.66	75.11
✓	✓	✓	**72.56**	**74.91**	**76.30**

Table 3: Ablation study on the contribution of each module. EAR: Energy-Aware Rebalancing; Syn: Synergistic Enhancement Projection; Orth: Conflict Orthogonalization Projection. Adding each component progressively improves performance, with all 3 combined achieving the best results.

more stable convergence. Together, these factors lead to superior generalization under multimodal imbalance.

Comparison with conflict modulation methods

Table 2 reports the performance comparison between our proposed RGM and 8 representative conflict-aware optimization methods across three datasets (KS, CREMA-D, UCF-51) and two backbones (CNN-based and Transformer-based). These baselines can be broadly divided into two categories: (1) general gradient conflict solvers such as GradNorm (Chen et al. 2018) and PCGrad (Yu et al. 2020), originally designed for multi-task learning and adapted to multimodal setups, and (2) multimodal-specific conflict modulation methods including MMCosine (Xu et al. 2023), GMD (Wang et al. 2024), ReconBoost (Hua et al. 2024), MMPareto (Wei and Hu 2024), CGGM (Guo et al. 2024), and BALGRAD (Kwon et al. 2025), which explicitly address modality conflicts through direction alignment, decomposition, or multi-objective optimization. Across all tasks and settings, our RGM achieves the superior accuracy. For example, under CNN-based backbones, RGM achieves 72.56% on KS and 76.3% on UCF-51, outperforming the strongest baselines ReconBoost and BALGRAD by +1.22% and +0.67%, respectively. Under Transformer-based settings, RGM remains competitive, reaching 62.47% on CREMA-D, surpassing BALGRAD by +0.82%.

These improvements stem from RGM's unified strategy that dynamically modulates gradient directions and magnitudes based on the degree of inter-modal agreement. Specifically, RGM encourages collaborative learning by aligning gradients of low-conflict modality pairs, while enforcing orthogonality for high-conflict pairs to preserve diversity and prevent destructive interference. Unlike prior methods that treat all conflicts equally or require handcrafted thresholds, RGM adapts modulation strength through a soft synergy–orthogonal blending, achieving finer-grained tradeoffs between fusion and separation.

Ablation Study

Table 3 presents the ablation study evaluating the impact of each proposed module. Specifically, EAR (Energy-Aware Rebalancing) mitigates training-stage imbalance by adaptively scaling gradient magnitudes; Syn (Synergistic Enhancement Projection) promotes positive interaction by projecting low-conflict gradients into a shared direction; Orth (Conflict Orthogonalization Projection) reduces interference by enforcing near-orthogonality for high-conflict modality gradients. Each module individually improves performance over the naive baseline, while their combinations lead to further gains. Notably, integrating all three achieves the best accuracy across all datasets. This highlights the complementary nature of our design: balancing modality strength (EAR), enhancing cooperation (Syn), and preserving diversity (Orth) are all essential for robust multimodal learning.

Qualitative Analysis

Training Dynamics on the Balance–nonConflict Plane. Fig. 3 visualizes the training dynamics of two-modality gradients projected onto the BCP, where the x-axis denotes the Balance factor $\bar{w}_{ij}$ and the y-axis denotes the nonConflict factor c_{ij}. Each point corresponds to a training sample at a specific iteration, color-coded by training step.

Fig. 3 (a) shows the vanilla Joint training baseline, where the training process gradually shifts toward imbalance (left-shifted $\bar{w}_{ij}$) without resolving conflicts (high variance in c_{ij}). In (b), PCGrad explicitly enforces alignment to increase gradient similarity. While this reduces conflicts, it inadvertently suppresses modality diversity, resulting in worse balance than (a). In (c), ARM targets only imbalance correction. Although it improves $\bar{w}_{ij}$, the conflict factor remains widely distributed around zero, indicating unresolved interference. Fig. 3 (d) shows CGGM, which jointly adjusts gradient magnitude and direction. However, its global projection strategy tends to over-align gradients, causing a loss of modality-specific representation.

In contrast, Fig. 3 (e) demonstrates the effectiveness of our proposed RGM. RGM not only drives the training trajectory toward a well-balanced region (higher $\bar{w}_{ij}$), but also adaptively handles conflict: gradients in low-conflict regions are synergistically aligned, while high-conflict pairs are orthogonally rotated. This results in a clear structure in the upper right quadrant, reflecting improved collaboration without homogenizing the expression of the modality.

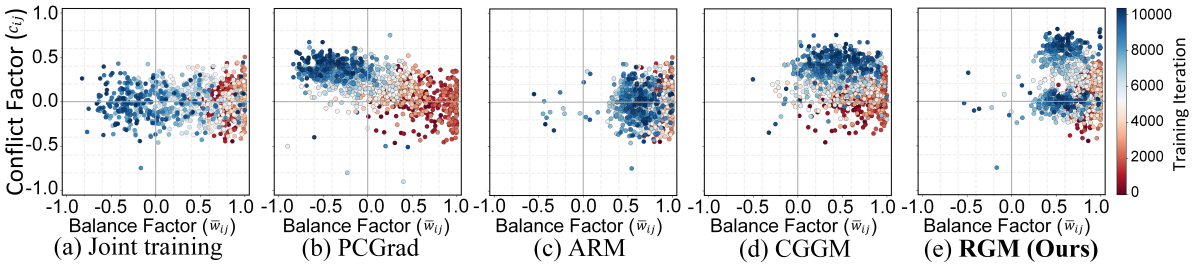

Figure 3: Visualization of training dynamics on the Balance-nonConflict Plane on Kinetics Sounds dataset. Each point represents a batch-sample's gradient interaction at a training step, where the x-axis is the Balance factor $\bar{w}_{ij}$ and the y-axis is the nonConflict factor c_{ij}. Color indicates training iteration (red = early, blue = later).

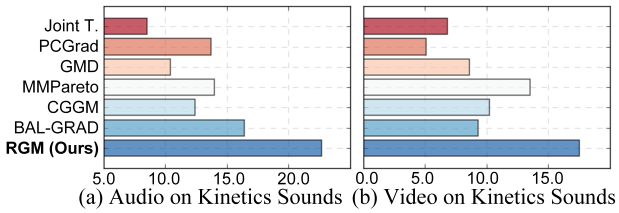

Figure 4: Modality-specific information on Kinetics Sounds.

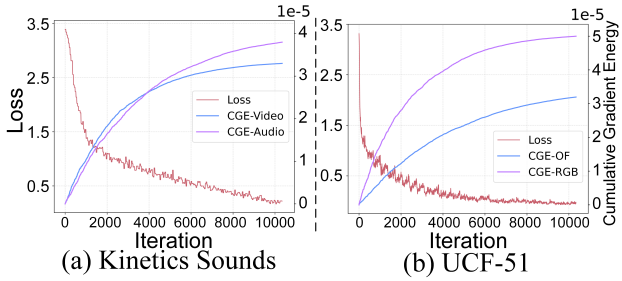

Figure 5: Curves of CGE and loss versus iteration.

Mutual Information Analysis To further assess the ability of each method to preserve modality-specific information, we conduct a mutual information analysis follow previous work (Liang et al. 2023; Hua et al. 2024), shown in Fig. 4. Specifically, we treat the performance drop when removing one modality (e.g., audio or video) as an empirical estimate of the specific task-relevant information that modality contributes. Formally, if x_i is retained and x_j is removed, the accuracy drop quantifies the added value of x_j in the multimodal setup. From the results on the Kinetics Sounds, we observe that RGM consistently yields the highest modality-specific information values for both audio and video modalities, significantly outperforming CGGM and other baselines. This confirms that our orthogonal projection mechanism prevents destructive interference and over-alignment between modalities, enabling each to retain their distinct semantic contributions throughout training.

Convergence analysis of CGE.

To verify the stability and effectiveness of our proposed Cumulative Gradient Energy (CGE) metric, we conduct a convergence analysis shown in Fig. 5. We track both the training loss and CGE curves of each modality across iterations. As illustrated in both (a) Kinetics Sounds and (b) UCF-51, the training loss rapidly decreases and stabilizes, while the CGE of each modality grows monotonically and gradually saturates as learning progresses. The convergence of CGE implies that the accumulated gradient signal from each modality becomes stable over time, aligning with the model's convergence. Importantly, CGE provides a dynamic and continuous measurement of each modality's learning progression, rather than relying on fixed snapshots. It captures both the magnitude and persistence of updates contributed by each modality, making it a more informative signal than instantaneous gradients or loss contributions. This property enables CGE to act as a modality-aware feedback signal, guiding reweighting or modulation strategies like EAR in a principled way. Moreover, the plateauing behavior confirms that CGE can be used to assess whether certain modalities dominate, collapse, or are under-trained during learning.

Conclusion

In this work, we present Reconcile Gradient Modulation (RGM), a unified framework that simultaneously addresses modality imbalance and conflict through adaptive modulation in the gradient space. RGM leverages a principled formulation, SynOrth Grad, to perform minimal-gradient surgery—enhancing synergy via co-linearity projection when modalities align and enforcing orthogonality to preserve diversity under conflict. To support informed modulation, we introduce Cumulative Gradient Energy (CGE) and construct the Balance–nonConflict Plane (BCP), which enables real-time monitoring and interpretation of the training dynamics with convergence guarantees. Extensive experiments verify the effectiveness and robustness of RGM. While this work focuses on conflicts at the representation level, our current formulation does not yet account for data-level conflicts, such as modality misalignment or semantic inconsistency. In future work, we plan to extend the RGM framework toward modeling and resolving such low-level cross-modal conflicts, thereby improving the adaptability of multimodal systems in more challenging real-world settings.

Acknowledgments

This work was sponsored by the National Natural ScienceFoundation of China (No.s 62222608, 62436002, 62476198), the Tianjin Natural Science Funds for Distinguished YoungScholar (No. 23JCJQJC00270), the Zhejiang ProvincialNatural Science Foundation of China (No. LD24F020004), and the Natural Science Foundation of Tianjin (No. 25JCYBJC00950).

References

Achille, A.; and Soatto, S. 2018. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50): 1–34.

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, 609–617.

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

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, 794–803. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Fisher, R. A. 1925. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, 5, 700–725. Cambridge University Press.

Gallo, I.; Calefati, A.; and Nawaz, S. 2017. Multimodal classification fusion in real-world scenarios. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 5, 36–41. IEEE.

Gao, X.; Cao, B.; Zhu, P.; Wang, N.; and Hu, Q. 2025. Asymmetric reinforcing against multi-modal representation bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, 16754–16762.

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.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, J.; Yan, S.; Cai, J.; Jiang, X.; Hu, Y.; and Xie, W. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.

Hu, P.; Li, X.; and Zhou, Y. 2022. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*.

Hua, C.; Xu, Q.; Bao, S.; Yang, Z.; and Huang, Q. 2024. Reconboost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*.

Huang, C.; Wei, Y.; Yang, Z.; and Hu, D. 2025. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25854–25863.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Imfeld, M.; Graldi, J.; Giordano, M.; Hofmann, T.; Anagnostidis, S.; and Singh, S. P. 2023. Transformer fusion with optimal transport. *arXiv preprint arXiv:2310.05719*.

Jiang, Q.-Y.; Chi, Z.; and Yang, Y. 2025. Interactive multimodal learning via flat gradient modification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 5489–5497.

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*.

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. *arXiv preprint arXiv:2503.13834*.

Li, H.; Li, X.; Hu, P.; Lei, Y.; Li, C.; and Zhou, Y. 2023a. Boosting Multi-modal Model Performance with Adaptive Gradient Modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22214–22224.

Li, Y.; Quan, R.; Zhu, L.; and Yang, Y. 2023b. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2604–2613.

Liang, P. P.; Deng, Z.; Ma, M. Q.; Zou, J. Y.; Morency, L.-P.; and Salakhutdinov, R. 2023. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36: 32971–32998.

Ma, X.; Chen, H.; and Deng, Y. 2025. Improving Multimodal Learning Balance and Sufficiency through Data Remixing. *arXiv preprint arXiv:2506.11550*.

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.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shen, M.; Huang, Y.; Yin, J.; Zou, H.; Rajan, D.; and See, S. 2023. Towards balanced active learning for multimodal classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3434–3445.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tzirakis, P.; Chen, J.; Zafeiriou, S.; and Schuller, B. 2021. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68: 46–53.

Wang, H.; Luo, S.; Hu, G.; and Zhang, J. 2024. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14, 15483–15491.

Wang, L.; Li, W.; Li, W.; and Van Gool, L. 2018. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1430–1439.

Wei, Y.; Feng, R.; Wang, Z.; and Hu, D. 2024a. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27338–27347.

Wei, Y.; and Hu, D. 2024. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*.

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

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

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

Yang, Y.; Wan, F.; Jiang, Q.-Y.; and Xu, Y. 2024. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems*, 37: 62108–62122.

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.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, 3987–3995. PMLR.

Zhang, H.; Hao, Y.; and Ngo, C.-W. 2021. Token shift transformer for video classification. In *Proceedings of the 29th acm international conference on multimedia*, 917–925.

Zhang, K.; He, L.; Jiang, X.; Lu, W.; Wang, D.; and Gao, X. 2025. CognitionCapturer: Decoding Visual Stimuli From Human EEG Signal With Multimodal Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14486–14493.

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

Zhou, Y.; Wang, X.; Chen, H.; Duan, X.; and Zhu, W. 2023. Intra-and inter-modal curriculum for multimodal learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3724–3735.