MSCR: Jointly Balancing Utilization Across Modalities and Discovering Synergistic Information

Xinyu Chen[1], **Liangjian Wen**[1,7†], **Jiang Duan**[1], **Dongkai Wang**[1], **Yong Dai**[3], **Jiayu Bai**[1], **Jianzhuang Liu**[2], **Zhen Tian**[5], **Guoping Qiu**[6], **Zhao Kang**[4]

[1]School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]X-Humanoid
[4]University of Electronic Science and Technology of China
[5]Guangdong Laboratory of Artificial Intelligence and Digital Economy
[6]Computer Science, University of Nottingham
[7]Artificial Intelligence and Digital Finance Key Laboratory of Sichuan Province

chenxinyu991@gmail.com, wlj6816@gmail.com

Abstract

In multimodal learning, modality imbalance often lets one modality dominate optimization and keeps the model from fully exploiting cross-modal interactions. Balance-oriented methods such as the Multimodal Competition Regularizer (MCR) perform well by alleviating this imbalance during training. These dynamic balancing methods, however, mainly adjust modality contributions and do not consider how imbalance affects the formation of synergistic information during fusion. Synergy is a major source of multimodal advantage, because it is predictive gain that cannot be recovered from unimodal evidence alone. We propose the Multimodal Synergistic Competition Regularizer (MSCR), a fusion method that jointly promotes balanced modality utilization and the discovery of synergistic information. MSCR has two complementary objectives. An anti-dominance objective penalizes large modality-perturbation sensitivities to regularize modality reliance, with the aim of leaving "representation room" for joint predictive cues. A synergy objective encourages the fused predictor to outperform an additive unimodal reference. An adaptive gate further concentrates the regularization on samples that need stronger collaborative modeling. Experiments on five benchmarks show that MSCR consistently improves performance over MCR and strengthens synergistic capability. It raises conditional recovery, the accuracy of the fused prediction on samples that the unimodal predictors misclassify, as well as the share of the model's output that InterSHAP attributes to cross-modal interactions. Its gain in conditional recovery persists when the fusion module is Transformer-based. Code is available at https://github.com/chenxinyu991-collab/MSCR.

1 Introduction

Multimodal learning integrates heterogeneous signals such as vision, audio, and language, and has achieved strong performance [25, 35, 54]. Recent work has extended it to high-resolution visual understanding [28] and knowledge-driven reasoning for web agents [11]. Its main advantage comes from combining complementary information. Jointly trained multimodal models, however,

†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

often suffer from modality imbalance: the model relies too heavily on a dominant modality, which hinders cross-modal fusion [38, 53, 17]. Prior work has explored several training strategies to alleviate this imbalance, including suppressing information from the dominant modality [49, 5, 39], monitoring modality contributions to adaptively modulate gradients [32, 29, 22, 20], and decomposing multimodal learning into alternating unimodal optimization [58], which preserves cross-modal interaction while reducing biased optimization dynamics.

The Multimodal Competition Regularizer (MCR) [21] was recently proposed to mitigate modality imbalance explicitly. MCR perturbs modality features to characterize each modality's contribution to the joint prediction. It treats each encoder as a greedy player that maximizes its own importance, possibly at the expense of the others. This competitive game balances modality utilization well, but it does not address the discovery of synergistic information [46, 3], that is, task information that is available only when the modalities are observed jointly. Synergistic gain comes from collaborative interactions between modalities, so forcing modalities to compete does not by itself push the model to find information that no single modality can provide. More broadly, earlier methods and MCR alike focus on adjusting modality contributions during training and pay little attention to how imbalance and purely competitive dynamics affect the formation of synergistic information [51, 41].

Synergistic information is a major source of multimodal advantage [41, 50, 42], because it provides predictive gain beyond unimodal evidence. Even a strong modality can be insufficient: identifying sarcasm, for example, requires interpreting facial expression and vocal intonation together [55, 56]. A model that over-relies on a dominant modality tends to exploit only the information that modality can supply on its own (its unique and redundant components) and to overlook synergistic information that requires joint observation.

We therefore introduce the Multimodal Synergistic Competition Regularizer (MSCR), which combines modality balancing with explicit synergy discovery. MSCR first limits the dominance of any single modality with an anti-dominance objective. It perturbs modality features in the latent space to measure how strongly the fused prediction depends on each modality, and penalizes the network for relying too heavily on any one of them. Discouraging excessive modality reliance may leave "representation room" for cross-modal interactions. This is the motivation for the constraint, but reducing perturbation sensitivity alone does not guarantee synergy. MSCR therefore adds a synergy objective based on an additive reference formed from independent unimodal predictions. Requiring the fused model to achieve a discriminative gain over this naive ensemble encourages the network to exploit synergistic information. Finally, to keep optimization stable and avoid over-regularization, an instance-level proficiency-aware gate tracks unimodal explanatory power during training and scales the MSCR regularizer for each sample, so that it acts mainly on difficult samples on which unimodal evidence is insufficient.

Our main contributions are as follows:

- We point out that modality imbalance suppresses the discovery of synergistic information, and we propose MSCR, a fusion method that jointly optimizes balanced modality utilization and synergistic gain.
- Motivated by an information-theoretic view, MSCR limits single-modality dominance in the fusion network and explicitly rewards joint predictions that outperform an additive unimodal reference. A proficiency-aware adaptive gate applies this regularization mainly to samples that unimodal evidence cannot resolve, which keeps optimization stable.
- Experiments show that MSCR consistently improves downstream performance and synergistic capability, as measured by conditional recovery and an independent InterSHAP [45] analysis, and that its advantage in conditional recovery persists with a Transformer-based fusion module.

2 Preliminaries: Multimodal Competition Regularizer

To mitigate modality imbalance, MCR [21] regulates the information contributed by each modality. Following MCR, we express the joint mutual information through an information-theoretic identity:

$$I(X_1, X_2; Y) = \underbrace{I(X_1; Y \mid X_2) + I(X_2; Y \mid X_1)}_{\text{Conditional Task-Relevant Information}} + \underbrace{I(X_1; X_2)}_{\text{Shared Info}} - \underbrace{I(X_1; X_2 \mid Y)}_{\text{Conditional Shared Info}} . \quad (1)$$

Based on this decomposition, MCR uses three regularization terms:

$$\mathcal{L}_{\mathrm{MCR}} = \mathcal{L}_{\mathrm{MIPD}} + \mathcal{L}_{\mathrm{Con}} + \mathcal{L}_{\mathrm{CEB}}. \tag{2}$$

Modality Contribution Learning ($\mathcal{L}_{\mathrm{MIPD}}$). MCR introduces the Mutual Information Perturbed Difference (MIPD) to characterize each modality's contribution to the joint prediction, motivated by the conditional mutual information $I(X_i; Y \mid X_j)$ for $i \neq j$. Under partial information decomposition (PID), this quantity includes both information unique to modality i and synergistic information available only through joint observation. MIPD uses changes in the prediction under modality perturbation as a practical proxy for contribution. Randomly permuting the latent representations of modality 1 across a mini-batch preserves their empirical marginal distribution but breaks their sample-wise pairing with modality 2 and the labels. Writing the perturbed modality as $\widetilde{X}_1$, MCR uses the negative Jensen–Shannon (JS) divergence [26] between the original and perturbed joint predictions as a stable optimization surrogate:

$$\mathcal{L}_{\mathrm{MIPD}_1} = -\mathbb{E}\left[\mathrm{JSD}\left(p(y \mid x_2, x_1), p(y \mid x_2, \widetilde{x}_1)\right)\right]. \tag{3}$$

$\mathcal{L}_{\mathrm{MIPD}_2}$ is defined analogously by perturbing X_2. Minimizing $\mathcal{L}_{\mathrm{MIPD}_i}$ maximizes this prediction discrepancy. Together with task supervision, it encourages the corresponding encoder to learn features that are used, rather than ignored, by the fusion network in the joint prediction.

Shared Information Alignment ($\mathcal{L}_{\mathrm{Con}}$). MCR uses a supervised cross-modal contrastive loss [19] to encourage shared class-relevant structure. Positive pairs include cross-modal representations from samples with the same class label, not just the two modalities of a single sample. The loss encourages greater agreement within these positive pairs than with other samples in the mini-batch.

Irrelevant Information Filtering ($\mathcal{L}_{\mathrm{CEB}}$). MCR interprets the conditional cross-modal dependence $I(X_1; X_2 \mid Y)$ as task-irrelevant shared information. It regularizes this dependence with a penalty inspired by the Conditional Entropy Bottleneck (CEB) [7]. A small reconstruction head h predicts the concatenated modality representations from the label Y. Minimizing the reconstruction error

$$\mathcal{L}_{\mathrm{CEB}} = \mathbb{E}\left[\left\|[f_1(x_1), f_2(x_2)] - h(y)\right\|^2\right], \tag{4}$$

encourages representations that can be predicted from the task label and discourages variation that the label does not explain. This penalty is a practical regularizer, not an exact measure of $I(X_1; X_2 \mid Y)$.

The Game of Competition. MCR treats each modality encoder as a player in a competitive game. Under the default greedy strategy ($k = -1$), encoder i minimizes its own $\mathcal{L}_{\mathrm{MIPD}_i}$ while maximizing the other modality's $\mathcal{L}_{\mathrm{MIPD}_j}$. The MIPD gradient used to update encoder parameters θ_i by gradient descent is proportional to

$$\nabla_{\theta_i} \mathcal{L}_{\mathrm{MIPD}_i} - \nabla_{\theta_i} \mathcal{L}_{\mathrm{MIPD}_j}, \qquad i, j \in \{1, 2\}, \quad i \neq j.$$

Each loss is a negative prediction discrepancy, so this update encourages each encoder to increase its own influence on the joint prediction while reducing the other's. The competition comes from these encoder-specific gradient rules, not from simply minimizing the sum of the two MIPD losses. It is intended to mitigate modality imbalance, but it neither explicitly isolates nor directly optimizes synergistic information.

3 Method

3.1 The General Framework

We consider a supervised multimodal learning task. For simplicity of exposition, we describe the method with two input modalities X_1 and X_2 and a target label Y; the framework extends to more than two modalities. During training, for a batch of B samples $\{(\mathbf{x}_1^{(b)}, \mathbf{x}_2^{(b)}, y^{(b)})\}_{b=1}^{B}$, two modality-specific encoders f_1 and f_2 map the inputs into latent representations $\mathbf{z}_1 = f_1(\mathbf{x}_1) \in \mathbb{R}^{B \times d_1}$ and $\mathbf{z}_2 = f_2(\mathbf{x}_2) \in \mathbb{R}^{B \times d_2}$. A fusion network f_c combines these representations into the joint prediction $p_{\mathrm{joint}} = f_c(\mathbf{z}_1, \mathbf{z}_2)$, and two auxiliary unimodal predictors q_1 and q_2 produce the modality-specific

outputs $p_1 = q_1(\mathbf{z}_1)$ and $p_2 = q_2(\mathbf{z}_2)$. For classification, p_{joint}, p_1, and p_2 are predictive distributions over the classes; Appendix A gives the regression variant.

Partial information decomposition (PID) [46, 3] splits the task-relevant information $I(X_1, X_2; Y)$ into redundant information (R) that either modality provides on its own, unique information (U_1, U_2) specific to one modality, and synergistic information (S) that emerges only when the modalities are combined. $\mathcal{L}_{\text{MCR}}$ regulates modality contributions through competition, but it does not explicitly separate or directly target synergistic information S, which depends on inter-modal interactions.

MSCR addresses this gap as a lightweight regularizer that acts only on the fusion network f_c and does not interfere with the modality-specific encoders f_1 and f_2. It constrains modality-perturbation sensitivity with the aim of mitigating excessive modality reliance and leaving "representation room" for joint predictive cues. A separate objective explicitly encourages joint predictive gain beyond unimodal aggregation. The overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{MCR}} + \mathcal{L}_{\text{MSCR}}, \tag{5}$$

where $\mathcal{L}_{\text{task}}$ is the supervised task loss computed on the joint prediction and on the two unimodal predictions, $\mathcal{L}_{\text{MCR}}$ balances modality contributions as described in Section 2, and $\mathcal{L}_{\text{MSCR}}$ is our synergy-enhancing regularizer:

$$\mathcal{L}_{\text{MSCR}} = g \cdot (\lambda_{\text{dom}} \mathcal{L}_{\text{dom}} + \lambda_{\text{syn}} \mathcal{L}_{\text{syn}}) . \tag{6}$$

The anti-dominance objective $\mathcal{L}_{\text{dom}}$ penalizes large modality-perturbation sensitivities in the fusion network, placing the most weight on the largest sensitivity. The synergy objective $\mathcal{L}_{\text{syn}}$ encourages joint predictive gain beyond unimodal aggregation. The proficiency-aware gate $g \in (0, 1)$ scales both objectives according to sample difficulty. The gate and both objectives are computed per sample, and $\mathcal{L}_{\text{MSCR}}$ is averaged over the mini-batch. MSCR is switched on after an initial training stage and applied within a fixed interval of epochs (Appendix A); Algorithm 1 in Appendix F gives the complete training procedure.

The regularizers operate on distinct parameter groups. For MCR, the fusion parameters are held fixed when computing MIPD, while gradients with respect to the fusion network's inputs are retained to update the modality encoders; the CEB term also updates its reconstruction head. For MSCR, modality representations are detached in all joint and perturbed predictions, and the unimodal reference and gate are also detached, so its gradients update only the fusion network. MIPD and $\mathcal{L}_{\text{dom}}$ therefore act on different parameters. These restrictions apply only to the regularizers; $\mathcal{L}_{\text{task}}$ trains the model through the standard gradient paths.

3.2 Anti-Dominance Objective

Modality dominance occurs when the fusion network relies too heavily on readily accessible cues from one modality and overlooks the others, a common problem in multimodal learning. MIPD encourages each encoder to learn features that contribute to the joint prediction, but their availability does not ensure that the fusion network uses them jointly. We therefore regularize the strength of the fused prediction's dependence on individual modalities. The aim is to mitigate excessive modality reliance and leave "representation room" for joint predictive cues.

We use the MIPD-based perturbation measure from Section 2, but reverse its optimization direction when updating the fusion network. Permuting modality i across a mini-batch gives the perturbed predictions $p_{\text{shuf1}} = f_c(\widetilde{\mathbf{z}}_1, \mathbf{z}_2)$ and $p_{\text{shuf2}} = f_c(\mathbf{z}_1, \widetilde{\mathbf{z}}_2)$. For each sample, the modality-perturbation sensitivity is

$$s_i = \text{JSD}\left(p_{\text{joint}} \parallel p_{\text{shuf}i}\right) = -\ell_{\text{MIPD}_i}, \qquad i \in \{1, 2\}, \tag{7}$$

where ℓ_{MIPD_i} is the per-sample loss term in Eq. (3). A larger s_i means that the prediction changes more when modality i is replaced with an unpaired representation. This sensitivity is a proxy for modality reliance, not a measure of purely unimodal information. MIPD encourages an encoder to increase its own contribution; our objective instead penalizes large sensitivities when updating the fusion network. We combine these sensitivities with a differentiable smooth maximum:

$$\mathcal{L}_{\text{dom}} = \text{smoothmax}(s_1, s_2) = \frac{1}{\tau_{\text{dom}}} \log\left(\exp(\tau_{\text{dom}} s_1) + \exp(\tau_{\text{dom}} s_2)\right), \tag{8}$$

where $\tau_{\text{dom}} \geq 0$ controls the sharpness of the approximation. Since $\partial \mathcal{L}_{\text{dom}} / \partial s_i = \exp(\tau_{\text{dom}} s_i) / \sum_j \exp(\tau_{\text{dom}} s_j)$, larger sensitivities receive more weight in the penalty. The most

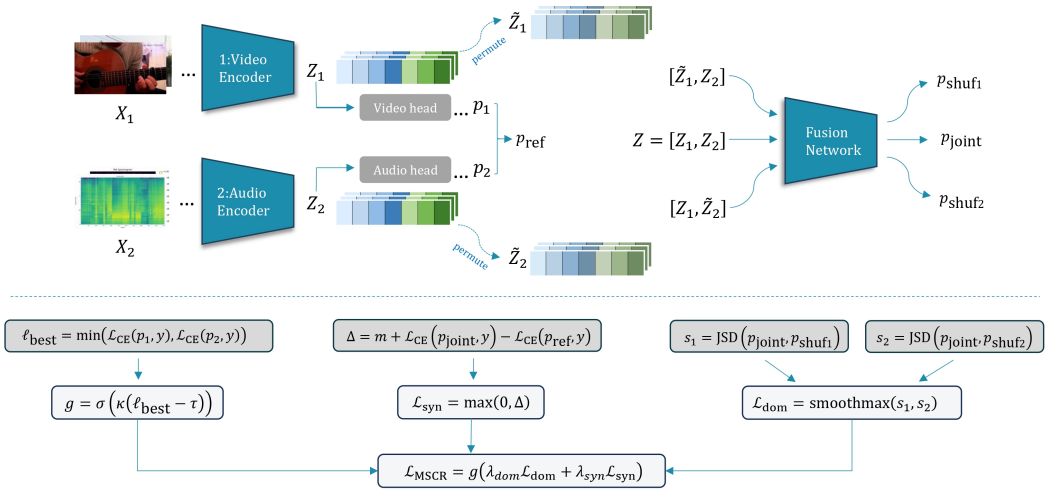


Figure 1: **Multimodal Synergistic Competition Regularizer** (MSCR)**:** MSCR suppresses modality dominance and encourages the discovery of synergistic information. Raw data (X_1, X_2) are encoded into representations (Z_1, Z_2). These and their permuted variants $(\tilde{Z}_1, \tilde{Z}_2)$ are fed into the fusion network to produce joint (p_{joint}) and shuffled $(p_{\text{shuf1}}, p_{\text{shuf2}})$ predictions, while the unimodal heads provide an additive reference (p_{ref}). The MSCR loss is built from three mechanisms: $\mathcal{L}_{\text{dom}}$ uses a smoothmax to penalize modality-perturbation sensitivities, placing the most weight on the largest sensitivity; $\mathcal{L}_{\text{syn}}$ requires the joint prediction to outperform the unimodal reference by a margin m, with Δ denoting the hinge argument; and a proficiency-aware gate (g) scales the regularizer according to the best unimodal loss (ℓ_{best}).

heavily penalized modality can therefore vary from sample to sample. We detach modality representations in this branch, so the objective changes how the fusion network uses the available features without directly updating the encoders.

Reducing perturbation sensitivity alone does not guarantee synergy, because permuting a modality can also disrupt useful cross-modal interactions. Leaving "representation room" is therefore a design motivation, not a guaranteed redistribution of information. We use $\mathcal{L}_{\text{dom}}$ as an auxiliary constraint together with the task loss, which encourages accurate predictions, and $\mathcal{L}_{\text{syn}}$, which encourages joint predictive gain beyond unimodal aggregation.

3.3 Synergy Objective

Even when no single modality dominates, the joint prediction may be no better than an aggregate of unimodal evidence and miss the predictive gains that arise only from combining modalities. MSCR therefore adds a margin-based synergy objective that rewards joint predictive gain beyond unimodal aggregation.

As a simple proxy for synergistic capability, we define an additive unimodal reference distribution $p_{\text{ref}} = \frac{1}{2}(p_1 + p_2)$, where p_1 and p_2 are the outputs of the auxiliary unimodal predictors q_1 and q_2 of the model being trained. This reference simply averages independent unimodal decisions. A fusion network that extracts meaningful synergistic information should substantially outperform this naive ensemble.

The task losses of the joint prediction and the additive reference are

$$\ell_{\text{joint}} = \mathcal{L}_{\text{CE}}(p_{\text{joint}}, y), \quad \text{and} \quad \ell_{\text{ref}} = \mathcal{L}_{\text{CE}}(p_{\text{ref}}, y), \tag{9}$$

where y is the ground-truth label and $\mathcal{L}_{\text{CE}}(p, y) = -\log p(y)$ is the cross-entropy loss. Thus ℓ_{ref} is the negative log of the average probability that the two unimodal predictors assign to the correct class. The synergy objective is

$$\mathcal{L}_{\text{syn}} = \max(0,\, m + \ell_{\text{joint}} - \ell_{\text{ref}}), \tag{10}$$

where $m > 0$ is a fixed margin. Minimizing $\mathcal{L}_{\text{syn}}$ drives the model toward $\ell_{\text{joint}} \leq \ell_{\text{ref}} - m$, so the joint predictor must reach a lower loss than the additive baseline. A sample that satisfies this condition no longer contributes to $\mathcal{L}_{\text{syn}}$, and because ℓ_{ref} is detached, the objective can be reduced

only by improving the joint prediction. The objectives complement each other: $\mathcal{L}_{\text{dom}}$ regularizes modality reliance through perturbation sensitivity, while $\mathcal{L}_{\text{syn}}$ explicitly encourages joint predictive gain beyond the unimodal reference.

3.4 Proficiency-Aware Adaptive Gating

Samples differ in how much regularization they need. When at least one unimodal predictor already classifies a sample correctly with high confidence, strong regularization on that sample may be unnecessary. When both unimodal predictions are ambiguous or wrong, the fusion model has to mine synergistic information to reduce task uncertainty. MSCR therefore uses an instance-level, proficiency-aware gate to adjust the strength of $\mathcal{L}_{\text{MSCR}}$.

For a given sample, let the unimodal task losses be $\ell_1 = \mathcal{L}_{\text{CE}}(p_1, y)$ and $\ell_2 = \mathcal{L}_{\text{CE}}(p_2, y)$. We measure the best unimodal proficiency by the minimum unimodal loss, $\ell_{\text{best}} = \min(\ell_1, \ell_2)$, and compute the gate $g \in (0, 1)$ with a smooth step function:

$$g = \sigma\big(\kappa(\ell_{\text{best}} - \tau)\big), \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function, and $\kappa > 0$ and $\tau > 0$ set the sharpness and the threshold of the gate (τ is distinct from the smoothmax temperature τ_{dom}). For classification, $\ell_{\text{best}} > \tau$ means that even the more confident unimodal predictor assigns a probability below $e^{-\tau}$ to the correct class, which is about 0.30 and 0.22 for the thresholds $\tau = 1.2$ and $\tau = 1.5$ used in our experiments.

The gate works as a soft switch. It opens ($g \to 1$) when unimodal evidence is insufficient ($\ell_{\text{best}} > \tau$), so both objectives concentrate on hard samples, and it closes ($g \to 0$) for easy samples that a single modality already handles ($\ell_{\text{best}} \ll \tau$), so that on these samples the fusion network is trained almost entirely by the task loss.

For the trimodal settings, p_{ref} averages the three unimodal predictions, the smooth maximum in Eq. (8) is taken over the three sensitivities, and ℓ_{best} is the smallest of the three unimodal losses.

4 Experiments

4.1 Main Results on Real-World Benchmarks

We first evaluate MSCR on real-world multimodal benchmarks. Following the evaluation protocol of MCR [21], we use CREMA-D [4] for emotion recognition, AVE [34] for event recognition, UCF101 [31] for action recognition, and CMU-MOSI [55] and CMU-MOSEI [56] for multimodal sentiment and emotion analysis. The datasets cover bimodal and trimodal settings as well as classification and regression-style tasks. We report prior multimodal learning methods under the same protocol and focus on the comparison between MSCR and the state-of-the-art baseline MCR.

Table 1 also includes the training strategies evaluated by MCR. Unimodals reports each modality trained on its own, and Ensemble combines these unimodal predictions without further training. Joint Training optimizes all modalities under a single multimodal loss, and Multi-Loss [37] adds unimodal task losses. Uni-Pre Frozen and Uni-Pre Finetuned start from pre-trained unimodal encoders that are kept frozen or fine-tuned jointly, respectively. The remaining baselines are the balancing methods reviewed in Section 5. In the table header, the second row gives the unimodal encoder and the third the modalities used (V: video, A: audio, T: text), and UCF denotes UCF101. Entries are the mean $\pm$ standard deviation over three folds (CREMA-D, UCF101) or three random seeds (AVE, MOSI, MOSEI), and the best result in each column is in bold. The encoders, fusion network, and training schedule follow MCR; Appendices A and B give the details.

As shown in Table 1, MSCR outperforms MCR in every evaluated setting. The advantage holds across task domains (audio-visual classification and multimodal sentiment analysis) and modality settings (bimodal and trimodal). The following subsections examine whether these gains come with stronger synergistic capability and how each component of MSCR contributes.

4.2 Synergy Analysis via Conditional Recovery

To assess synergistic capability beyond balanced modality utilization, we analyze conditional recovery rates. The idea is to examine the multimodal model on "hard" cases in which both unimodal predictors

Table 1: Comparison of MSCR with prior multimodal training methods on five real-world benchmarks, covering bimodal and trimodal settings. Following MCR, the regression-style tasks (CMU-MOSI and CMU-MOSEI) are evaluated by binary accuracy, and the other datasets are standard classification tasks. For a fair comparison, we follow the evaluation protocol of MCR. MCR results are reproduced by us, while the remaining baseline results are taken from that work.

	CREMA-D		AVE		UCF	MOSI		MOSEI	
	ResNet	Conformer	ResNet	Conformer	ResNet	Transformer			
Method	V-A	V-A	V-A	V-A	V-A	V-T	V-A-T	V-T	V-A-T
Unimodals	V: 55.4$_{\pm 3.0}$ A: 60.6$_{\pm 2.3}$	V: 69.4$_{\pm 2.0}$ A: 76.0$_{\pm 2.6}$	V: 45.7$_{\pm 1.6}$ A: 62.6$_{\pm 0.9}$	V: 75.5$_{\pm 1.2}$ A: 76.5$_{\pm 2.4}$	V: 38.5$_{\pm 0.9}$ A: 30.3$_{\pm 1.5}$	V: 54.1$_{\pm 3.7}$ A: 53.7$_{\pm 0.6}$ T: 72.1$_{\pm 3.3}$		V: 64.8$_{\pm 0.2}$ A: 64.4$_{\pm 0.2}$ T: 78.9$_{\pm 1.7}$	
Ensemble	71.7$_{\pm 2.2}$	84.6$_{\pm 1.0}$	70.5$_{\pm 0.2}$	88.4$_{\pm 2.2}$	52.8$_{\pm 0.5}$	70.5$_{\pm 2.1}$	67.2$_{\pm 1.5}$	78.4$_{\pm 0.7}$	77.2$_{\pm 0.6}$
Joint Training	62.6$_{\pm 5.8}$	74.6$_{\pm 2.2}$	66.7$_{\pm 1.5}$	82.2$_{\pm 0.8}$	47.7$_{\pm 1.5}$	73.0$_{\pm 1.3}$	73.6$_{\pm 1.3}$	80.5$_{\pm 0.2}$	80.8$_{\pm 0.3}$
Multi-Loss [37]	69.2$_{\pm 1.8}$	82.6$_{\pm 0.9}$	70.1$_{\pm 0.9}$	86.3$_{\pm 1.1}$	51.1$_{\pm 1.8}$	72.1$_{\pm 0.4}$	73.6$_{\pm 2.9}$	80.0$_{\pm 0.7}$	80.2$_{\pm 0.5}$
Uni-Pre Frozen	72.4$_{\pm 1.8}$	85.0$_{\pm 1.8}$	72.2$_{\pm 0.3}$	87.2$_{\pm 2.4}$	53.0$_{\pm 0.9}$	73.3$_{\pm 1.8}$	72.7$_{\pm 1.6}$	79.9$_{\pm 0.5}$	79.8$_{\pm 0.3}$
Uni-Pre Finetuned	73.3$_{\pm 1.8}$	82.4$_{\pm 2.0}$	72.5$_{\pm 1.3}$	86.5$_{\pm 0.8}$	53.5$_{\pm 1.3}$	73.1$_{\pm 2.3}$	73.7$_{\pm 0.7}$	80.3$_{\pm 0.4}$	80.3$_{\pm 0.2}$
MSLR [53]	56.5$_{\pm 2.4}$	77.1$_{\pm 2.4}$	67.3$_{\pm 2.2}$	81.0$_{\pm 1.4}$	50.9$_{\pm 3.9}$	×	×	×	×
MMCosine [52]	59.3$_{\pm 1.5}$	74.0$_{\pm 0.3}$	65.0$_{\pm 1.4}$	83.8$_{\pm 0.8}$	47.3$_{\pm 4.1}$	×	×	×	×
OGM [29]	65.6$_{\pm 3.8}$	82.4$_{\pm 1.0}$	67.3$_{\pm 0.6}$	79.7$_{\pm 1.4}$	51.8$_{\pm 1.9}$	73.9$_{\pm 1.1}$	⊗	79.7$_{\pm 0.6}$	⊗
AGM [22]	69.3$_{\pm 1.4}$	78.5$_{\pm 1.6}$	68.4$_{\pm 1.1}$	85.3$_{\pm 0.5}$	51.0$_{\pm 1.6}$	74.0$_{\pm 1.3}$	73.9$_{\pm 1.9}$	79.3$_{\pm 0.4}$	80.2$_{\pm 0.3}$
MLB [20]	71.9$_{\pm 2.2}$	85.2$_{\pm 0.9}$	71.6$_{\pm 0.2}$	86.7$_{\pm 0.3}$	52.2$_{\pm 1.7}$	72.4$_{\pm 1.7}$	74.2$_{\pm 1.7}$	80.1$_{\pm 0.5}$	80.5$_{\pm 0.4}$
ReconBoost [16]	69.0$_{\pm 2.4}$	84.8$_{\pm 1.8}$	68.4$_{\pm 1.7}$	86.1$_{\pm 0.7}$	50.2$_{\pm 4.0}$	×	×	×	×
MMPareto [39]	69.0$_{\pm 2.5}$	83.8$_{\pm 0.8}$	73.0$_{\pm 1.3}$	87.4$_{\pm 1.3}$	51.4$_{\pm 2.2}$	73.4$_{\pm 1.0}$	73.7$_{\pm 0.6}$	79.3$_{\pm 0.6}$	79.5$_{\pm 0.8}$
D&R [40]	70.6$_{\pm 1.3}$	85.0$_{\pm 0.4}$	72.3$_{\pm 1.5}$	**91.0$_{\pm 0.7}$**	49.3$_{\pm 1.0}$	×	×	×	×
MCR	74.5$_{\pm 0.8}$	85.1$_{\pm 0.5}$	73.5$_{\pm 0.0}$	87.4$_{\pm 0.2}$	52.9$_{\pm 2.4}$	75.3$_{\pm 0.7}$	78.6$_{\pm 0.2}$	80.8$_{\pm 0.4}$	81.1$_{\pm 0.4}$
MSCR	**75.7$_{\pm 1.1}$**	**86.4$_{\pm 0.9}$**	**74.5$_{\pm 0.2}$**	89.3$_{\pm 0.3}$	**54.4$_{\pm 2.4}$**	**76.8$_{\pm 0.8}$**	**80.1$_{\pm 0.5}$**	**81.3$_{\pm 0.5}$**	**81.8$_{\pm 0.6}$**

× method not applicable to regression tasks;
⊗ method not applicable to trimodal inputs

Table 2: Conditional recovery rates (Eq. (12)) of MCR and MSCR, i.e., the accuracy of the multimodal model on samples that the unimodal predictors misclassify. Higher values are better.

	CREMA-D		AVE		UCF	MOSI		MOSEI	
	ResNet	Conformer	ResNet	Conformer	ResNet	Transformer			
Method	V-A	V-A	V-A	V-A	V-A	V-T	V-A-T	V-T	V-A-T
MCR	10.13%	5.92%	5.50%	6.93%	10.28%	4.85%	10.87%	7.12%	6.71%
MSCR	**18.35%**	**9.98%**	**7.07%**	**12.95%**	**15.45%**	**34.35%**	**17.44%**	**12.50%**	**13.69%**

fail to provide sufficient evidence. Let $\hat{y}_{\text{joint}}$, $\hat{y}_1$, and $\hat{y}_2$ be the predictions of the fusion network and of the two auxiliary unimodal predictors of the same trained model, and let $\mathcal{W}$ be the set of test samples on which both $\hat{y}_1$ and $\hat{y}_2$ are wrong. The conditional recovery rate is the accuracy of the joint predictor on $\mathcal{W}$:

$$\text{Recovery} = \frac{\left|\{a \in \mathcal{W} : \hat{y}_{\text{joint}}^{(a)} = y^{(a)}\}\right|}{|\mathcal{W}|}. \tag{12}$$

For MOSI and MOSEI, correctness follows the binary accuracy criterion. In the trimodal settings, $\mathcal{W}$ contains the samples on which all three unimodal predictions are wrong. Unlike standard task metrics, this rate counts joint predictions that succeed where neither unimodal predictor does, which gives a behavioral proxy for synergistic capability. In Figures 2 and 3, it is the share of the Both Wrong column that falls in the Correct row.

Figure 2 shows the error comparison on CREMA-D with ResNet [12] and Conformer [10] unimodal backbones, and Table 2 reports the conditional recovery rates for all settings. The recovery rate of MSCR is higher than that of MCR on every dataset and backbone. The largest gain appears on MOSI in the V-T setting (from 4.85% to 34.35%). These results suggest that MSCR improves synergistic capability at the behavioral level: the joint predictor makes better use of latent cues that the unimodal predictors do not fully exploit. This pattern is consistent with the proficiency-aware gate, which puts more weight on samples with high unimodal prediction losses.

4.3 Synergy Analysis via Cross-Modal Interaction Assessment

As a complement to conditional recovery, we measure model-level cross-modal interaction with Global InterSHAP [45]. Built on the Shapley Interaction Index, it separates modality-specific contributions to the model output from interaction contributions. For sample a, let $\phi_{ij}^{(a)}$ $(i \neq j)$ be the Shapley interaction index of modalities i and j, and let $\phi_{ii}^{(a)}$ be the main effect of modality i,

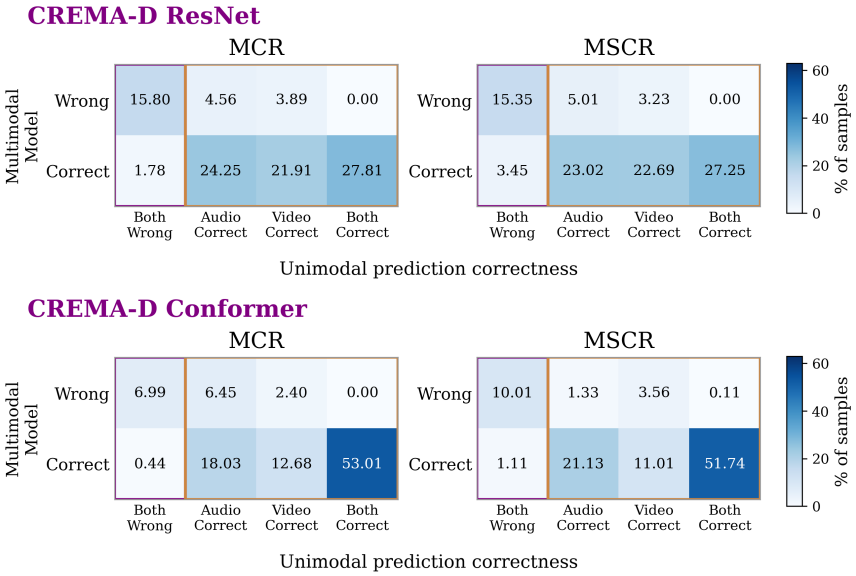


Figure 2: Error comparison on CREMA-D with ResNet and Conformer unimodal backbones. Each matrix shows multimodal prediction performance conditioned on unimodal prediction correctness, with cells giving the percentage of samples in each category. Columns indicate whether both unimodal predictors are wrong, only audio is correct, only video is correct, or both are correct; rows indicate whether the multimodal model is correct. Results on the other datasets are given in Appendix C.

Table 3: Global InterSHAP scores (%, mean $\pm$ standard deviation) under identical architectures. Higher scores indicate greater relative interaction contributions.

Dataset / Backbone	MCR	MSCR
CREMA-D / Conformer	5.22 ± 1.33	$\mathbf{10.31 \pm 1.89}$
AVE / Conformer	4.74 ± 0.93	$\mathbf{9.16 \pm 1.28}$
UCF / ResNet	12.10 ± 0.67	$\mathbf{18.21 \pm 1.12}$

i.e., its Shapley value minus its interactions with the other modalities. Over N evaluated samples, $\Phi_{ij} = \left| \frac{1}{N} \sum_{a=1}^{N} \phi_{ij}^{(a)} \right|$ is the absolute dataset-averaged contribution, so diagonal terms represent modality-specific effects and off-diagonal terms represent interactions. The global score is

$$\text{Global InterSHAP} = \frac{\sum_{i \neq j} \Phi_{ij}}{\sum_{i,j} \Phi_{ij}}. \tag{13}$$

The score lies in $[0, 1]$, is reported in percent, and measures the share of the model's output behavior that comes from cross-modal interactions. It requires neither auxiliary unimodal predictors nor ground-truth labels, so it complements the behavioral evidence from conditional recovery.

As shown in Table 3, MSCR increases the relative contribution of cross-modal interactions in all three settings under identical architectures, in line with its higher conditional recovery rates.

4.4 Synergy under Highly Interactive Fusion

The lower recovery rates of MCR might simply reflect the limited cross-modal interaction capacity of the default fusion module, which follows MCR and combines the modality representations through a single linear layer (Appendix A). To test this, we replace it with a Transformer-based fusion module implemented with two Conformer blocks, which apply self-attention to the concatenated multimodal token sequence. We then measure conditional recovery (Eq. (12)) for MCR and MSCR.

Table 4 shows that, even with self-attention across the concatenated multimodal tokens, MCR achieves lower conditional recovery rates than MSCR across all evaluated settings. On CREMA-D, MCR recovers 15.66% of the both-wrong samples with ResNet encoders and 11.82% with Conformer

Table 4: Conditional recovery rates under Transformer-based fusion architectures. Higher values indicate a greater ability to recover correct joint predictions when both unimodal predictors fail.

	CREMA-D		AVE	UCF
	ResNet	Conformer	ResNet	ResNet
Method	V-A	V-A	V-A	V-A
MCR	15.66%	11.82%	11.11%	16.41%
MSCR	**21.39%**	**25.66%**	**15.13%**	**22.63%**

encoders, whereas MSCR recovers 21.39% and 25.66%. A more expressive fusion module therefore does not close the gap in conditional recovery between MCR and MSCR.

4.5 Ablation Study

To understand the contribution of each component, we run an ablation on representative settings from the main experiments and compare four variants: the baseline without either objective (which reduces to MCR), the variant with only the anti-dominance objective $\mathcal{L}_{\text{dom}}$, the variant with only the synergy objective $\mathcal{L}_{\text{syn}}$, and the full MSCR with both objectives.

Table 5: Ablation study of the anti-dominance objective $\mathcal{L}_{\text{dom}}$ and the synergy objective $\mathcal{L}_{\text{syn}}$ across multiple datasets, backbones, and modality settings.

$\mathcal{L}_{\text{dom}}$	$\mathcal{L}_{\text{syn}}$	CREMA-D Conformer	AVE Conformer	UCF ResNet	MOSI V-T	MOSI V-A-T	MOSEI V-T	MOSEI V-A-T
		85.1±0.5	87.4±0.2	52.9±2.4	75.3±0.7	78.6±0.2	80.8±0.4	81.1±0.4
✓		85.9±0.6	88.8±0.4	53.6±2.2	75.9±0.7	79.7±0.5	81.1±0.3	81.4±0.4
	✓	86.2±0.5	88.3±0.2	53.5±2.3	76.1±0.9	79.2±0.6	80.9±0.4	81.2±0.2
✓	✓	**86.4±0.9**	**89.3±0.3**	**54.4±2.4**	**76.8±0.8**	**80.1±0.5**	**81.3±0.5**	**81.8±0.6**

As shown in Table 5, adding either $\mathcal{L}_{\text{dom}}$ or $\mathcal{L}_{\text{syn}}$ alone improves over the baseline in every evaluated setting, and optimizing both together gives the best results, so the two objectives are complementary.

5 Related Work

Imbalanced multimodal learning. Multimodal joint training often suffers from modality laziness [38, 47, 32, 5], in which the model strongly prefers dominant modalities and suppresses the learning of the others. Balance-oriented strategies address this by intervening in the optimization dynamics. One category uses dynamic modulation: OGM [29], AGM [22], and MMPareto [39] adjust gradients or loss weights to penalize the dominant modality [37, 52, 20], and MSLR [53] adaptively scales learning rates according to performance discrepancies. Another category decouples training: D&R [40], MLA [58], and ReconBoost [16] isolate unimodal updates to avoid interference between modalities. MCR [21] recently introduced an information-decomposition framework to mitigate dominance, and IBML [30] addresses modality imbalance by adaptively balancing the retention of complementary information across modalities. These approaches mainly aim to equalize the contributions of individual modalities and do not require the fusion network to produce collaborative gains, so a balanced model may still fail to exploit synergistic information.

Information-theoretic multimodal learning. Information bottleneck methods study the trade-off between compressing and preserving information [14]. MIGE [44] directly estimates mutual information gradients for representation learning, and MVEB [43] learns minimal sufficient representations by jointly maximizing the agreement between augmented views and the differential entropy of the embedding distribution. For multiview clustering, S2WIB [27] combines self-supervised view weighting with complementary information and cluster-structure consistency, and PIB [15] couples feature subspace learning and clustering through iterative information propagation. Beyond preserving information, one can ask whether predictive information is available from individual modalities or requires joint observation. PID [46, 3] formalizes this distinction through redundant, unique, and synergistic components. Recent studies [50, 51] argue that the main benefit of multimodal learning comes from synergistic information, that is, cues that emerge only from joint interaction and cannot be recovered from unimodal evidence alone [23]. To capture such interactions, FactorCL [23] and

CoMM [6] model interaction patterns with multimodal augmentations and contrastive objectives, and masking strategies have been used to uncover hidden synergistic semantics [41]. Our work targets supervised multimodal fusion: we introduce a fusion-level regularizer that encourages joint predictive gain beyond unimodal aggregation and can complement existing modality-balancing methods.

6 Conclusion

We proposed MSCR, a fusion-level regularizer that addresses modality imbalance and explicitly encourages the discovery of synergistic information. MSCR acts only on the fusion bottleneck and does not interfere with the unimodal encoders. It penalizes large perturbation sensitivities to regularize modality reliance and encourages the joint prediction to outperform an additive unimodal reference. A proficiency-aware gate adjusts the regularization strength for each sample. Experiments on five benchmarks show that MSCR consistently improves downstream performance over MCR and increases synergistic capability, as measured by conditional recovery and InterSHAP. A limitation is that the synergy objective uses an additive unimodal reference as a tractable approximation, which may not capture every form of multimodal interaction. A natural next step is a more exact synergy measure with theoretical bounds.

Acknowledgments

This work was supported by the Sichuan Provincial Innovation Group Project (Grant No. 2024NSFTD0054), the Major Science and Technology Special Project of the Sichuan Provincial Department of Science and Technology (Grant No. 2024ZDZX0002), Fundamental Research Funds for the Central Universities (JBK202511081), the National Natural Science Foundation of China (No. U24A20323, No. 62502397). We thank anonymous reviewers for their insightful feedback and Konstantinos Kontras, the first author of MCR [21], for his contributions to open sourcing the code.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in neural information processing systems*, volume 33, pages 12449–12460, 2020.

[3] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[4] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[5] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. *arXiv preprint arXiv:2305.01233*, 2023.

[6] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

[7] Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020.

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[9] Lucas Goncalves, Seong-Gyun Leem, Wei-Cheng Lin, Berrak Sisman, and Carlos Busso. Versatile audio-visual learning for handling single and multi modalities in emotion regression and classification tasks. *arXiv preprint arXiv:2305.07216*, 2023.

[10] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[11] Yuhan Guo, Cong Guo, Aiwen Sun, Hongliang He, Xinyu Yang, Yue Lu, Yingji Zhang, Xuntao Guo, Dong Zhang, Jianzhuang Liu, Jiang Duan, Yijia Xiao, Liangjian Wen, Hai-Ming Xu, and Yong Dai. Web-CogReasoner: Towards multimodal knowledge-induced cognitive reasoning for web agents. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[14] Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5325–5344, 2024.

[15] Shizhe Hu, Zenglin Shi, Xiaoqiang Yan, Zhengzheng Lou, and Yangdong Ye. Multiview clustering with propagating information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9915–9929, 2024.

[16] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

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

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020.

[20] Konstantinos Kontras, Christos Chatzichristos, Matthew Blaschko, and Maarten De Vos. Improving multimodal learning with multi-loss gradient modulation. *arXiv preprint arXiv:2405.07930*, 2024.

[21] Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Pu Liang, Matthew B. Blaschko, and Maarten De Vos. Balancing multimodal training through game-theoretic regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[23] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[24] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie (Yufan) Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[25] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[26] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[27] Zhengzheng Lou, Yucong Wu, Ke Zhang, Chaoyang Zhang, Bo Ji, and Shizhe Hu. Structure-enhanced self-supervised weighted information bottleneck for multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 37(8):3727–3740, 2026.

[28] Zhongkuan Mao, Wenzhuo Zhao, Xianjie Liu, Yidong Wang, Zhao Gao, Ronghao Xian, Yao Jiang, Yi Zhang, Liangjian Wen, and Keren Fu. Pay more attention to text in high-resolution MLLMs. *arXiv preprint arXiv:2609.23495*, 2026.

[29] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 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, 2022.

[30] Yang Qin, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep information-balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9384–9396, 2026.

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[32] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021.

[33] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[34] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*, pages 247–263, 2018.

[35] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569. Association for Computational Linguistics, 2019.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

[37] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. CentralNet: A multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision Workshops*, 2018.

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

[39] Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

[40] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86. Springer, 2024.

[41] Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, zhao kang, Jun Wang, Zenglin Xu, and Jiang Duan. InfMasking: Unleashing synergistic information by contrastive multimodal interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[42] Liangjian Wen, Linjie Li, Jiang Duan, Yong Dai, Jianzhuang Liu, and Zhao Kang. Dependency, compression, and synergy: A unified information-theoretic view of multimodal learning. *arXiv preprint arXiv:2609.14421*, 2026.

[43] Liangjian Wen, Xiasi Wang, Jianzhuang Liu, and Zenglin Xu. MVEB: Self-supervised learning with multi-view entropy bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6097–6108, 2024.

[44] Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, and Zenglin Xu. Mutual information gradient estimation for representation learning. In *International Conference on Learning Representations*, 2020.

[45] Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21501–21509, 2025.

[46] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[47] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and N. A. Moubayed. On modality bias in the TVQA dataset. *ArXiv*, abs/2012.10210, 2020.

[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 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, 2020.

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

[50] Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L. Ballard, Tianlong Chen, and Qi Long. I^2MoE: Interpretable multimodal interaction-aware mixture-of-experts. In *Forty-second International Conference on Machine Learning*, 2025.

[51] Lixin Xiu, Xufang Luo, and Hideki Nakayama. A comprehensive information-decomposition analysis of large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[52] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. MMCosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023.

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

[54] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[55] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[56] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246. Association for Computational Linguistics, 2018.

[57] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[58] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the claims, contributions, and scope of the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses the limitation of using an additive unimodal reference as a tractable approximation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper is primarily an empirical study of a new regularization method. We do not claim new theoretical theorems or formal proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed descriptions of the model architecture and loss functions are provided in method section and experiment section. The MSCR hyperparameters are listed in Appendix A, and the remaining settings follow MCR.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our code is publicly available at `https://github.com/chenxinyu991-collab/MSCR`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided detailed dataset preprocessing information and experimental settings in the Appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All main results in Table 1 and Table 5 report the mean and standard deviation. Depending on the standard evaluation protocol of each specific dataset, these statistics are calculated over either 3 independent runs with different random seeds or 3-fold cross-validation.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide detailed information on the computer resources used in Appendix A.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, our work fully complies with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: MSCR is a general-purpose training regularizer for multimodal fusion and is not tied to a specific task or application, so the work has no direct societal impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our research does not involve high-risk generative models or the release of sensitive private data.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets, pretrained models, and code we use are cited in the paper, and Appendix E summarizes their license information.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve new crowdsourcing or human subject experiments.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This research does not involve human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were only used for language polishing and formatting, which does not impact the core methodology.

Appendix

A Experimental details

This section gives the details needed to reproduce our experiments. For all datasets we follow the evaluation protocol of MCR [21] and use the same data splits, modality-specific encoders, fusion backbones, optimizer settings, learning-rate schedules, batch sizes, and training epochs. The checkpoint with the best validation performance is evaluated on the test set. For CREMA-D [4] and UCF [31] we use a three-fold evaluation protocol and report the mean and standard deviation across folds. For AVE [34], MOSI [55], and MOSEI [56] we report the mean and standard deviation over three independent runs with different random seeds.

For the audio-visual classification datasets (CREMA-D, AVE, and UCF) we report classification accuracy. For the multimodal sentiment analysis datasets (MOSI and MOSEI) we follow the regression protocol and metrics of MCR. Experiments with pre-trained unimodal encoders use the same initialization and fine-tuning strategy as MCR. In the regression tasks on CMU-MOSI and CMU-MOSEI, the unimodal and joint predictors produce scalar outputs, so we compute modality sensitivities with squared error instead of JSD and use MSE for the joint, reference, and unimodal losses in the synergy objective and the adaptive gate. The reference is the mean of the unimodal outputs.

Following MCR, our default fusion module sums modality-specific linear projections with a shared bias, which is equivalent to applying a single linear prediction layer to the concatenated modality representations. For the fusion replacement experiment in Section 4.4, the Transformer-based fusion module consists of two Conformer blocks with input and hidden dimensions of 512 and eight attention heads, and applies self-attention to the concatenated multimodal token sequence.

Table 6: Hyperparameters for MSCR. λ_{dom} and λ_{syn} denote the weights of the anti-dominance objective and the synergy objective. m is the margin in the synergy objective. τ_{dom} is the smoothmax temperature. κ and τ are the sharpness and threshold of the adaptive modulation gate.

dataset	λ_{dom}	λ_{syn}	m	τ_{dom}	κ	τ
CREMA-D ResNet V-A	0.02	0.03	0.05	10.0	5.0	1.5
CREMA-D Conformer V-A	0.05	0.03	0.05	10.0	5.0	1.2
AVE ResNet V-A	0.02	0.03	0.05	10.0	5.0	1.5
AVE Conformer V-A	0.05	0.03	0.05	10.0	5.0	1.2
UCF ResNet V-A	0.02	0.03	0.05	10.0	5.0	1.5
MOSI Transformer V-T	0.05	0.03	0.04	10.0	5.0	1.2
MOSI Transformer V-A-T	0.05	0.03	0.04	10.0	5.0	1.2
MOSEI Transformer V-T	0.05	0.03	0.04	10.0	5.0	1.2
MOSEI Transformer V-A-T	0.05	0.03	0.04	10.0	5.0	1.2

Table 6 lists the MSCR-specific hyperparameters (anti-dominance weight, synergy weight, margin, smoothmax temperature, gate sharpness, and gate threshold) for each setting. λ_{syn}, τ_{dom}, and κ are shared across all settings; λ_{dom}, m, and τ differ only between backbone families. MSCR is activated during epochs $[T_s, T_e] = [10, 30]$ for models with ResNet or Transformer encoders and $[5, 15]$ for models with Conformer encoders (Algorithm 1), with both endpoints included. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory. A single training run takes approximately 0.5–1.5 hours on CMU-MOSI and CMU-MOSEI, and around 5 hours on CREMA-D, AVE, and UCF, depending on the experimental setting. The code, configuration files, and running scripts are publicly available at `https://github.com/chenxinyu991-collab/MSCR`.

B Dataset and backbone details

B.1 Datasets

Following the experimental protocols of MCR [21], we describe each dataset and its preprocessing below.

CREMA-D [4] is an audio-visual emotion recognition dataset of video clips from 91 actors, each expressing six emotions: Happy, Sad, Anger, Fear, Disgust, and Neutral. Following the preprocessing in [38, 29, 22], video frames are sampled at 1 fps (3 consecutive frames) and audio at 22 kHz (3 seconds). Log-Mel spectrograms are computed via STFT with a window size of 512 and a step size of 353. For the stronger models we follow [9] and use 16 kHz audio and pre-computed facial features. Our data split also follows [9], so no actor appears in more than one of the training, validation, and test sets.

AVE [34] contains 4,143 videos in 28 event categories, each with temporally labeled audio-visual events of at least 2 seconds. Following [38], we sample 4 video frames at 1 fps and resample audio at 16 kHz with the same STFT settings as CREMA-D. We use the official predefined splits and report the standard deviation across three random seeds on the test set.

UCF101 [31] is a large-scale action recognition dataset. Following the multimodal protocol in [38], we use the subset of 51 action categories that contain both audio and video streams. Preprocessing is the same as for AVE. We use the standard 3-fold split of the original dataset and report the standard deviation across these folds.

CMU-MOSI [55] and **CMU-MOSEI** [56] are multimodal sentiment analysis benchmarks in which each segment is annotated with a sentiment score from -3 to 3. Standard metrics for these datasets [35] include 7-class accuracy, binary accuracy, F1 score, mean absolute error, and correlation; following MCR, we report binary accuracy. We use the aligned versions of both datasets provided by MultiBench [24]. Results are the mean and standard deviation across three random seeds on the test set.

B.2 Models: Backbone Unimodal Encoders

Following previous work [22, 29, 52, 38], our standard experiments use ResNet-18 [12] as the unimodal encoder for both video and audio in CREMA-D, AVE, and UCF101. These encoders are randomly initialized and use adaptive pooling to handle different input dimensions. For CMU-MOSI and CMU-MOSEI we use a 5-layer Transformer [36] backbone, following [25, 23].

To test whether MSCR remains effective with encoders of different capacity, we also evaluate stronger, pre-trained unimodal encoders.

CREMA-D Conformer setting. Following [9], we use the first 12 layers of Wav2Vec2 [2, 48], initialized with self-supervised pre-trained weights for speech recognition and fine-tuned during training. For video, MTCNN [57] extracts face bounding boxes and a frozen EfficientNet-B2 [33] serves as the feature extractor. A 5-layer Conformer [10] trained from scratch further processes the audio and facial features. The full model has about 183M parameters. We refer to this configuration as the "Conformer" setting.

AVE Conformer setting. We use a similar architecture for AVE. The visual branch is a ViViT model [1] pre-trained on Kinetics [18], and the audio branch is HuBERT [13] pre-trained on AudioSet [8]. A 5-layer Conformer [10] processes both representations, for about 232M parameters in total. With these encoders, the pre-trained knowledge covers non-speech audio events and complex temporal patterns in video.

C Supplementary synergy analysis

This section gives the remaining error comparison matrices for the conditional recovery analysis in Section 4.2, covering AVE [34] with ResNet and Conformer [10] backbones, MOSI [55] and MOSEI [56] in the V-T setting, and UCF [31] with the ResNet backbone. The error comparison matrices are computed on the test split from a single run for each setting, using seed 42 for datasets evaluated across random seeds and fold 0 for datasets evaluated across cross-validation folds. In contrast, Global InterSHAP scores are reported as mean $\pm$ standard deviation across random seeds or cross-validation folds, following the evaluation protocol of each dataset. Each matrix breaks down the joint predictions by the correctness of the individual unimodal predictions. The "Both Wrong" cases, in which both unimodal predictors fail, provide behavioral evidence of the joint predictor's synergistic capability.

Consistent with the main text, MSCR recovers a larger share of the Both Wrong samples than MCR [21] in all five settings of Figure 3.

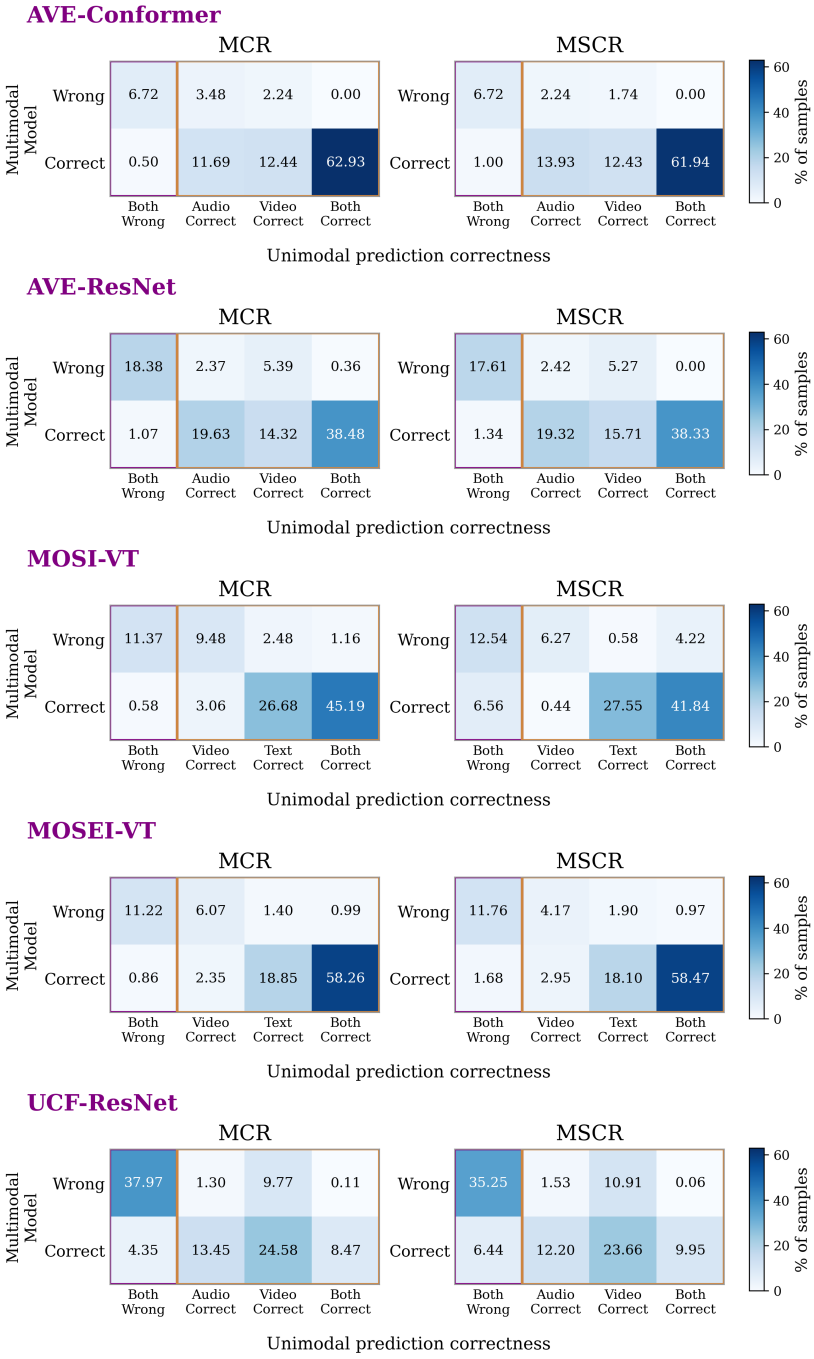


Figure 3: Error comparison matrices for additional datasets and backbones. Each matrix groups the multimodal predictions by the correctness of the corresponding unimodal predictions. Compared with MCR, MSCR achieves higher conditional recovery in the "Both Wrong" cases (unimodal predictors fail but the multimodal model is correct), which indicates a stronger ability to extract synergistic information when unimodal evidence is insufficient.

D Independence and Compatibility of MSCR

Our main experiments combine MSCR with MCR [21], but the regularizer does not depend on MCR-specific objectives. To test this, we remove all MCR regularization terms and keep only the supervised task loss and MSCR (*MSCR-only*). We also add MSCR to another modality-balancing method, OGM [29]. Within each setting, all variants share the same modality encoders and fusion architecture. Table 7 shows that MSCR-only performs comparably to MCR and that combining MSCR with MCR further improves accuracy. Adding MSCR to OGM improves performance in all three settings.

Table 7: Independence and compatibility of MSCR. MSCR-only removes all MCR-specific regularization terms. Results are accuracy (%), reported as mean ± standard deviation. OGM results are taken from [21]; all other results are obtained from our own experiments.

Method	CREMA-D Conformer	AVE Conformer	UCF ResNet
OGM	82.4 ± 1.0	79.7 ± 1.4	51.8 ± 1.9
OGM + MSCR	83.5 ± 1.2	81.1 ± 1.0	52.9 ± 1.7
MCR	85.1 ± 0.5	87.4 ± 0.2	52.9 ± 2.4
MSCR-only (w/o MCR)	85.3 ± 0.7	86.8 ± 0.4	52.7 ± 2.1
MCR + MSCR (full)	$\mathbf{86.4 \pm 0.9}$	$\mathbf{89.3 \pm 0.3}$	$\mathbf{54.4 \pm 2.4}$

E Third-Party Assets and Licenses

We summarize the licensing information for third-party assets below. Resource links and available license notices will be documented in our code repository.

Datasets. CREMA-D is distributed under the Open Database License (ODbL 1.0), with individual database contents licensed under the Database Contents License (DbCL 1.0). For CMU-MOSI and CMU-MOSEI, we use the preprocessed features provided through MultiBench. MultiBench and the CMU Multimodal SDK provide MIT-licensed code; this software license should be distinguished from rights in the underlying source videos. The official AVE and UCF101 pages do not state a dataset license.

Pretrained models and code. The model cards for `facebook/wav2vec2-large-robust`, `ALM/hubert-base-audioset`, and `google/vivit-b-16x2-kinetics400` specify Apache 2.0, CC BY-NC-SA 4.0, and MIT licenses, respectively. The Conformer implementation in the MCR codebase includes an Apache 2.0 license notice. We acknowledge the MCR and InterSHAP implementations. The README of the MCR repository indicates the MIT license, and the InterSHAP repository does not state a license.

F Pseudo-code

Algorithm 1 gives the training procedure of MSCR for bimodal classification. For each mini-batch, the model extracts modality-specific representations and computes unimodal and fused predictions. Within the MSCR activation interval, it builds modality-perturbed fused predictions to estimate modality dominance, forms an additive unimodal reference for synergy enhancement, and uses an adaptive gate to scale the regularization. Outside the activation interval, training uses the supervised task loss and MCR regularization.

Algorithm 1 Multimodal Training with MSCR

Input: Training dataset $\mathcal{D}$; modality encoders f_1, f_2; unimodal predictors q_1, q_2; fusion network f_c; MSCR weights $\lambda_{\text{dom}}, \lambda_{\text{syn}}$; margin m; smoothmax temperature τ_{dom}; gate parameters κ, τ; activation interval $[T_s, T_e]$.
Notation: $\text{sg}(\cdot)$ stops gradients; f_c^{fix} uses fixed fusion parameters while retaining gradients with respect to its inputs.
for each training epoch t **do**
  for each mini-batch $\mathcal{B} = \{(\mathbf{x}_1, \mathbf{x}_2, y)\}$ **do**
    Extract representations $\mathbf{z}_1 = f_1(\mathbf{x}_1)$, $\mathbf{z}_2 = f_2(\mathbf{x}_2)$.
    Compute unimodal predictions $p_1 = q_1(\mathbf{z}_1)$, $p_2 = q_2(\mathbf{z}_2)$, and joint prediction $p_{\text{joint}} = f_c(\mathbf{z}_1, \mathbf{z}_2)$.
    Compute the supervised task loss $\mathcal{L}_{\text{task}}$ using $p_{\text{joint}}, p_1, p_2$.
    MCR branch: regularize modality encoders.
    Compute $\mathcal{L}_{\text{MCR}}$ according to Eq. (2), using f_c^{fix} for MIPD and retaining gradients through $\mathbf{z}_1, \mathbf{z}_2$.
    MCR regularizers do not update f_c; the CEB term also updates its reconstruction head.
    if $T_s \le t \le T_e$ **then**
      MSCR branch: regularize the fusion network.
      Detach representations: $\bar{\mathbf{z}}_1 = \text{sg}(\mathbf{z}_1)$, $\bar{\mathbf{z}}_2 = \text{sg}(\mathbf{z}_2)$.
      Compute $\bar{p}_{\text{joint}} = f_c(\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2)$.
      Independently permute $\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2$ along the batch dimension to obtain $\tilde{\bar{\mathbf{z}}}_1, \tilde{\bar{\mathbf{z}}}_2$.
      Compute $p_{\text{shuf1}} = f_c(\tilde{\bar{\mathbf{z}}}_1, \bar{\mathbf{z}}_2)$, $p_{\text{shuf2}} = f_c(\bar{\mathbf{z}}_1, \tilde{\bar{\mathbf{z}}}_2)$.
      Compute per-sample sensitivities $s_i = \text{JSD}(\bar{p}_{\text{joint}} \| p_{\text{shuf}i})$, $i \in \{1, 2\}$.
      Compute $\ell_{\text{dom}} = \frac{1}{\tau_{\text{dom}}} \log(e^{\tau_{\text{dom}} s_1} + e^{\tau_{\text{dom}} s_2})$.
      Construct the detached reference $p_{\text{ref}} = \text{sg}\big(\frac{p_1 + p_2}{2}\big)$.
      Compute $\ell_{\text{joint}} = \mathcal{L}_{\text{CE}}(\bar{p}_{\text{joint}}, y)$, $\ell_{\text{ref}} = \mathcal{L}_{\text{CE}}(p_{\text{ref}}, y)$.
      Compute the per-sample synergy loss $\ell_{\text{syn}} = \max(0, m + \ell_{\text{joint}} - \ell_{\text{ref}})$.
      Compute detached unimodal losses $\ell_i = \text{sg}(\mathcal{L}_{\text{CE}}(p_i, y))$, $i \in \{1, 2\}$.
      Compute the per-sample gate $g = \sigma(\kappa[\min(\ell_1, \ell_2) - \tau])$.
      Aggregate over the mini-batch: $\mathcal{L}_{\text{MSCR}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} g^{(b)} \left(\lambda_{\text{dom}} \ell_{\text{dom}}^{(b)} + \lambda_{\text{syn}} \ell_{\text{syn}}^{(b)} \right)$.
    else
      Set $\mathcal{L}_{\text{MSCR}} = 0$.
    end if
    Set $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{MCR}} + \mathcal{L}_{\text{MSCR}}$.
    Backpropagate $\mathcal{L}_{\text{total}}$ through the designated gradient paths and update parameters.
  end for
end for