Balancing Multimodal Training Through Game-Theoretic Regularization

Konstantinos Kontras[1*] **Thomas Strypsteen**[1] **Christos Chatzichristos**[1]

Paul Pu Liang[3] **Matthew Blaschko**[1] **Maarten De Vos**[1,2]

[1]Department of Electrical Engineering, KU Leuven, Leuven, Belgium
[2]Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[3]MIT Media Lab and EECS, Cambridge, MA, USA

Abstract

Multimodal learning holds promise for richer information extraction by capturing dependencies across data sources. Yet, current training methods often underperform due to modality competition, a phenomenon where modalities contend for training resources leaving some underoptimized. This raises a pivotal question: how can we address training imbalances, ensure adequate optimization across all modalities, and achieve consistent performance improvements as we transition from unimodal to multimodal data? This paper proposes the Multimodal Competition Regularizer (MCR), inspired by a mutual information (MI) decomposition designed to prevent the adverse effects of competition in multimodal training. Our key contributions are: 1) A game-theoretic framework that adaptively balances modality contributions by encouraging each to maximize its informative role in the final prediction 2) Refining lower and upper bounds for each MI term to enhance the extraction of both task-relevant unique and shared information across modalities. 3) Proposing latent space permutations for conditional MI estimation, significantly improving computational efficiency. MCR outperforms all previously suggested training strategies and simple baseline, clearly demonstrating that training modalities jointly leads to important performance gains on both synthetic and large real-world datasets. We release our code and models at https://github.com/kkontras/MCR.

1 Introduction

Exploiting multimodal data has made significant progress, with advances in generalizable representations and larger datasets enabling solutions to previously unattainable tasks [28, 33, 30, 39, 45, 44, 46, 52, 55, 66]. However, studies indicate that jointly trained multimodal data is often utilized suboptimally, underperforming compared to ensembles of unimodal models, jointly trained modalities, or even the best single modality [56, 64]. The expectation that adding a new modality should improve performance, assuming independent errors and above-chance predictive power [16], is frequently contradicted in practice.

Huang et al. [20] attribute this issue to modality competition, where one modality quickly minimizes training error, misdirecting and suppressing the learning of others. To counteract this effect, monitoring each modality's contribution during training and applying corrective measures is crucial. To this

*Correspondence to: konstantinos.kontras@kuleuven.be

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

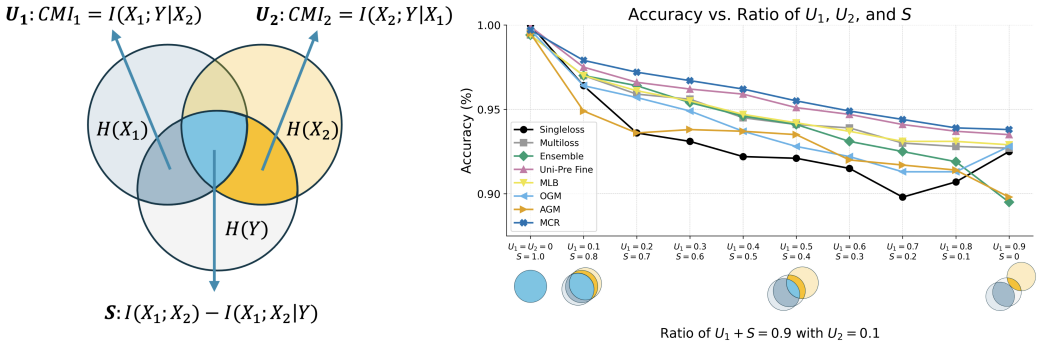


Figure 1: (Left) Illustration of the conditional mutual information (CMI) terms, $\text{CMI}_1 : I(X_1; Y \mid X_2)$ and $\text{CMI}_2 : I(X_2; Y \mid X_1)$, representing the unique contributions (U_1, U_2) of each modality. The shared task-relevant information (S) is defined as $I(X_1; X_2) - I(X_1; X_2 \mid Y)$. (Right) Accuracy on a synthetic dataset designed to induce **multimodal competition**. We vary the ratio of unique information from modality 1 (U_1) to shared information (S), while keeping the contribution of modality 2 (U_2) constant. As the imbalance increases (moving right on the x-axis), the performance of most methods drops. The standard Joint Training (Singleloss) approach shows a steep decline, highlighting its vulnerability to modality competition where one modality dominates and suppresses the other. In contrast, our method, MCR, demonstrates greater robustness by maintaining the highest accuracy and exhibiting the slowest performance degradation. See Section 4.1 for more details.

end, several balancing strategies have been proposed [5, 6, 9, 10, 21, 27, 29, 42, 43, 56, 61, 64, 57, 59, 69, 19, 58]. Some ignore a modality's contribution beyond its independent unimodal performance, while others address this by measuring output differences under input perturbation, but at the cost of increased sensitivity to these perturbations and significant computational overhead. Moreover, it is crucial to examine whether enhancing one modality's influence on the output does not come at the expense of others, as this could undermine overall performance.

Given these challenges, how can we efficiently regularize multimodal competition to ensure balanced and effective learning across modalities?

This paper introduces a loss function encouraging the exploration of task-relevant information across modalities, the MULTIMODAL COMPETITION REGULARIZER (MCR). The approach incorporates the following key contributions:

1. **MI Bounds:** We decompose joint mutual information into task-relevant shared and unique components, using refined lower and upper bounds to promote informative signals and suppress noise.
2. **Game-Theoretic Modality Balancing:** We frame modality interaction through a game-theoretic framework, allowing each modality to adjust its contribution throughout training.
3. **Efficient CMI Estimation:** We introduce latent-space perturbations for low-cost conditional MI estimation, avoiding repeated full-model passes.

We extensively evaluate MCR on synthetic datasets and several established real-world multimodal benchmarks, including action recognition on AVE [51] and UCF [47], emotion recognition on CREMA-D [4], human sentiment on CMU-MOSI [65], human emotions on CMU-MOSEI [67], and egocentric action recognition on Something-Something [14]. Our results demonstrate that MCR outperforms all previous methods and simple baselines across various datasets and models, improving multimodal supervised training.

2 Problem Analysis and Related Work

Consider a dataset of N independent and identically distributed (i.i.d.) datapoints sampled from a distribution $\mathcal{D}$, where each datapoint has M modalities $X = (X_1, \ldots, X_M)$ and a target Y_t. Our goal is to learn a parameterized function $f : X; \theta \to Y_t$, where θ denotes all the model's learnable parameters. The unimodal encoder for each modality is defined as $f_m : X_m; \theta_m \to Z_m$, encoding input X_m into a latent representation Z_m. The fusion network $f_c : [Z_1, \ldots, Z_M]; \theta_c \to Y_t$ predicts Y_t from the latent representations, as do the unimodal task heads $f_{c_m} : Z_m; \theta_{c_m} \to Y_m$. Model

families are defined as, unimodal models for $m = [1, .., M]$ modalities:

$$\mathcal{F}_{u_m} : f_{u_m}\left(X_m; \theta_m, \theta_{c_m}\right) = f_{c_m}\left(f_m\left(X_m; \theta_m\right); \theta_{c_m}\right), \tag{1}$$

and for multimodal models:

$$\mathcal{F} : f(X; \theta) = f_c\left(\left[f_1\left(X_1; \theta_1\right), .., f_M\left(X_M; \theta_M\right)\right]; \theta_c\right). \tag{2}$$

For simplicity, we continue our analysis with $M = 2$, focusing on models with two modalities.

2.1 The limitation of supervised multimodal training

In supervised learning, the goal is to learn representations $Z_1 = f_1(X_1; \theta_1)$ and $Z_2 = f_2(X_2; \theta_2)$ when fused via $f_c([Z_1, Z_2])$, yield accurate predictions. This is achieved by minimizing the task loss or, equivalently, by maximizing the MI between the fused representation and the target:

$$\underset{\substack{Z_1 := f_1(X_1; \theta_1),\\ Z_2 := f_2(X_2; \theta_2)}}{\arg\max} \quad I(f_c([Z_1, Z_2]); Y_t). \tag{3}$$

During training, models often over-rely on the stronger or more accessible modality, limiting the contribution of others. This leads to mutual information being dominated by one modality, e.g., $I(f_c([Z_1, Z_2]); Y) \approx I(Z_1; Y)$ with $I(Z_2; Y \mid Z_1) \approx 0$, indicating that Z_2 adds little once Z_1 is learned. See Appendix A.1 for an illustrative experiment and Appendix A.2 for a formal definition of the resulting generalization gap.

This kind of imbalance is well-known in single-modality learning, where dominant features can overshadow others, harming generalization. Regularization techniques like l_1 / l_2 penalties and dropout promote balanced feature use [40, 48], but their adaptation to multimodal settings is nontrivial. For example, applying modality-specific dropout [62] offers limited benefits [42]. The core challenge remains: how to effectively regulate interaction and competition between modalities.

2.2 Related Work

Prior research has explored various strategies for multimodal learning, ranging from simple unimodal and ensemble-based approaches to more sophisticated methods for balancing modality contributions. Unimodal training optimizes each modality separately, while ensemble methods combine unimodal predictions without additional training. Joint training optimizes all modalities under a single-loss objective but does not explicitly ensure sufficient training for each modality. To address this, Multi-Loss [54] introduces additional unimodal task losses, and MMCosine [63] equalizes modality influence by standardizing features and weights. Pre-trained unimodal encoders are often used, either with frozen weights (Uni-Pre Frozen) or fine-tuned jointly (Uni-Pre Finetuned). Other adaptive strategies include MSLR [64], which adjusts learning rates based on unimodal validation performance, OGM [42], which modulates gradients by comparing unimodal performance across modalities, and MLB [27], which combines unimodal task losses and modulates gradients from both unimodal and multimodal objectives, MMPareto [57] that mitigates gradient conflicts between modalities by equalizing the contribution of unimodal and multimodal gradient and D&R [59] suggest a new strategy where modalities that overfit get their part of the network partially reweighted with the initial weights of the training.

Most of these methods assume distributional independence and measure modality contributions through unimodal performance, which can be a limited indicator, missing cases where modality correlation is crucial. Other approaches estimate influence based on prediction differences after perturbations [29, 21, 10]. AGM [29] uses zero-masking Shapley values directly optimizing them as unimodal predictors, Wei et al. [58] use a permutation-based Shapley values and resampling of the training set to affect the training, while other methods address similar problems by introducing perturbations such as Gaussian noise [10] or task-specific augmentations [21, 32]. However, perturbation-based approaches increase the network's sensitivity to the chosen perturbations and hinder scalability due to their higher computational demands.

A line of work keeps unimodal training as the primary strategy. MLA [69] uses a shared task head and dynamic weighted summation during validation, while ReconBoost [19] alternates unimodal updates with agreement and diversity regularization before finetuning the ensemble. However, these approaches avoid multimodal training in the earlier steps to mitigate conflicts, yet overlook the potential benefits of direct multimodal interactions in those steps.

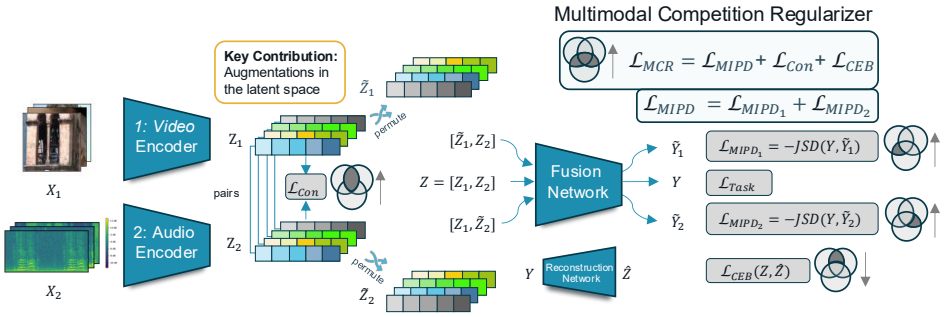


Figure 2: **Multimodal Competition Regularizer** (MCR)**:** The diagram illustrates the MCR framework, which mitigates modality competition in multimodal learning. Raw data (X_1 and X_2) are encoded into latent representations (Z_1 and Z_2), which are then permuted to create $\tilde{Z}_1$ and $\tilde{Z}_2$ and the paired combinations. These combinations are passed through the Fusion Network to produce predicted outputs (Y, $\tilde{Y}_1$, $\tilde{Y}_2$). The comparison between predictions reveals each modality's contribution. For example, if $Y \approx \tilde{Y}_1$, it shows that X_1 has little impact, and the model relies on X_2. The MCR loss includes three components: $\mathcal{L}_{\text{MIPD}}$ maximize the Jensen-Shannon divergence (JSD) between task output and permuted modality predictions. $\mathcal{L}_{\text{Con}}$ aligns modality representations, while $\mathcal{L}_{\text{CEB}}$ penalizes task-irrelevant information by reconstructing back to the latent space.

3 Multimodal Competition Regularizer

Multimodal competition arises when a model trained on multiple modalities prioritizes one, leading to over-reliance and reducing the contribution of others. This imbalance limits the model's ability to fully utilize all available information. In this section, we introduce $\mathcal{L}_{\text{MCR}}$, a set of loss components designed to address multimodal competition. Each component of the loss is motivated by the following MI decomposition:

$$I(X_1; X_2; Y) = \underbrace{I(X_1; Y \mid X_2) + I(X_2; Y \mid X_1)}_{\substack{\text{Task-Relevant Unique Information} \\ \text{of each modality} \sim \mathcal{L}_{\text{MIPD}}}} + \underbrace{I(X_1; X_2)}_{\substack{\text{Shared Information} \\ \mathcal{L}_{\text{Con}}}} - \underbrace{I(X_1; X_2 \mid Y)}_{\substack{\text{Task-Irrelevant} \\ \text{Shared Information} \sim \mathcal{L}_{\text{CEB}}}}. \tag{4}$$

This decomposition is illustrated by the Venn diagram in Figure 1. The CMIs $I(X_1; Y \mid X_2)$ and $I(X_2; Y \mid X_1)$ capture modality-specific information for predicting the target. Maximizing them with the Mutual Information Perturbed Difference (MIPD) loss, $\mathcal{L}_{\text{MIPD}}$, which assesses each modality's contribution via output variations under input perturbations (elaborated in Sec. 3.3) and encourages the extraction of modality-specific, task-relevant features. The third term, $I(X_1; X_2)$, quantifies shared information between modalities. Maximizing it with a contrastive loss, $\mathcal{L}_{\text{Con}}$, aligns representations and leverages their shared information effectively [22, 41, 46]. The final term, $I(X_1; X_2 \mid Y)$, represents task-irrelevant shared information. Penalizing it with the conditional entropy bottleneck (CEB) [8] and the corresponding loss $\mathcal{L}_{\text{CEB}}$ to filter out irrelevant information, focusing the model on features relevant to the downstream task. Each term has a corresponding loss, as illustrated in Figure 2, forming the regularizer with three key losses:

$$\mathcal{L}_{\text{MCR}} = \mathcal{L}_{\text{MIPD}} + \mathcal{L}_{\text{Con}} + \mathcal{L}_{\text{CEB}} \tag{5}$$

3.1 Approximating MI Terms

$\mathbf{I(X_1; Y \mid X_2)}$: To approximate each CMI and capture the unique contribution of each modality, the MIPD serves as a surrogate function, measuring how input perturbations affect the model's output. By comparing predictions with and without these perturbations, MIPD estimates how much information each modality provides. If a modality is crucial, altering its input should significantly change the output, revealing its importance.

Estimating the CMI directly through $I(X_1; Y \mid X_2) = H(Y \mid X_2) - H(Y \mid X_1, X_2)$ is typically intractable. Instead we use the MIPD as a lower bound, defined as:

$$\text{MIPD}(X_1; Y \mid X_2) = I(X_1; Y \mid X_2) - I(\tilde{X}_1; Y \mid X_2) \leq I(X_1; Y \mid X_2), \tag{6}$$

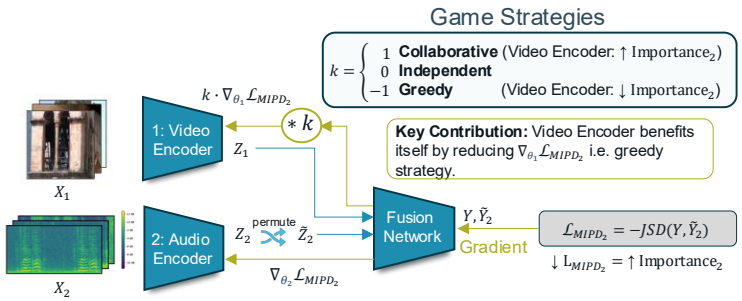


Figure 3: This figure illustrates a key aspect of our training process, showing how competition strategies between modalities are applied. The gradient multiplier adjusts the video encoder's response to Audio Importance (Importance$_2$). When $k = 1$, the video encoder enhances Importance$_2$; at $k = 0$, it remains neutral, and at $k = -1$, it competes by reducing Importance$_2$ to prioritize its own (Importance$_1$). This reflects the principle that increasing the importance of one modality can reduce the importance of the other.

where the perturbed version of modality X_1 is denoted as $\tilde{X}_1$. Interpreting MI via entropy, each CMI can be expressed as the difference of the log probabilities with and without the perturbations:

$$\begin{aligned} \mathrm{MIPD}(X_1; Y \mid X_2) &= H(Y \mid X_2, \tilde{X}_1) - H(Y \mid X_2, X_1) \\ &= \mathop{\mathbb{E}}_{\substack{y \sim p(y) \\ x_1, x_2 \sim p(x_1, x_2, y)}} \left[- \mathop{\mathbb{E}}_{\tilde{x}_1 \sim p(x_1)} [\log p(y \mid x_2, \tilde{x}_1)] + \log p(y \mid x_2, x_1) \right]. \end{aligned} \tag{7}$$

Instead of this log-likelihood ratio, we use the symmetrically bounded Jensen-Shannon divergence (JSD) [34] to prevent training instabilities, leading to the following:

$$\mathcal{L}_{\mathrm{MIPD}_1} = -\mathrm{MIPD}(X_1; Y \mid X_2) = - \mathop{\mathbb{E}}_{\substack{y \sim p(y) \\ x_1, x_2 \sim p(x_1, x_2, y) \\ \tilde{x}_1 \sim p(x_1)}} [\mathrm{JSD}(p(y \mid x_2, x_1), p(y \mid x_2, \tilde{x}_1))]. \tag{8}$$

Similarly, $\mathcal{L}_{\mathrm{MIPD}_2}$ can be computed symmetrically.

$\mathbf{I(X_1; X_2)}$: The next MI term measures how much information the two modalities share, capturing the common patterns between the modalities and aligning the representations of these shared aspects. We exploit the available label information employing the supervised contrastive loss $\mathcal{L}_{\mathrm{Con}}$ [25]:

$$\mathcal{L}_{\mathrm{Con}} = \mathop{\mathbb{E}}_{\substack{x_1, y \sim p(x_1, y) \\ x_2^+ \sim p(x_2 \mid y) \\ x_2^- \sim p(x_2 \mid \neg y)^2}} \left[\log \frac{\psi(x_1, x_2^+)}{\sum_k \psi(x_1, x_{2_k}^-)} \right], \tag{9}$$

where ψ is the critic function, which, in our case, is the exponential dot product. Minimizing the $\mathcal{L}_{\mathrm{Con}}$, maximizes a lower bound on both the MI between the two modalities and the CMI terms:

$$I(X_2; Y \mid X_1) + I(X_1; Y \mid X_2) + 2I(X_2; X_1) \geq \log N - \mathcal{L}_{\mathrm{Con}}^{Opt}. \tag{10}$$

As N increases, the bound becomes tighter, while the bound is not affected by the number of positive samples (same class datapoints). More details are provided in Appendix B.

$\mathbf{I(X_1; X_2 \mid Y)}$: The final term captures irrelevant shared information between modalities, and minimizing an upper bound on this ensures the model retains only task-relevant content. For this purpose, we exploit the idea of Conditional Entropy Bottleneck (CEB) L_{CEB} [7], targeting superfluous information in multimodal representations via a reconstruction loss. A small reconstruction head, $h : Y; \theta_h \rightarrow Z = (Z_1, Z_2)$, predicts back the latent space, effectively filtering out irrelevant content:

$$\mathcal{L}_{\mathrm{CEB}} = \mathop{\mathbb{E}}_{x_1, x_2, y \sim p(x_1, x_2, y)} \| [f_1(x_1), f_2(x_2)] - h(y; \theta_h) \|^2 \tag{11}$$

The exact derivation of this loss term can be found in Appendix A.6. Penalizing irrelevant information has been shown to enhance calibration and robustness [8], but it must be carefully evaluated, as it can introduce constraints that may hinder overall performance.

[2] $\neg y = \{y' \in \mathcal{Y} \mid y' \neq y\}$, where $\mathcal{Y}$ is the set of target labels.

3.2 The Game of Multimodal Fusion

We adopt a game-theoretic approach to balance the terms of the proposed $\mathcal{L}_{\text{MIPD}}$. The key idea is that increasing one modality's importance (e.g., via MIPD_1) can inherently reduce the other's (e.g., MIPD_2). Thus, an underutilized encoder i (with parameters θ_i) can boost its relevance both by minimizing $\mathcal{L}_{\text{MIPD}_i}$ and by maximizing $\mathcal{L}_{\text{MIPD}_{\neg i}}$. This twofold strategy helps prevent suppression of weaker modalities. We frame $\mathcal{L}_{\text{MIPD}}$ as a game where each encoder (player) selects a strategy, minimize, maximize, or ignore. Figure 3 illustrates how the video modality, via a hyperparameter k, can choose to assist, ignore, or diminish the audio modality. Each encoder applies this logic selectively as formalized below:

$$\nabla_{\theta_1}\mathcal{L}_{\text{MIPD}} = \lambda_M \left(\nabla_{\theta_1}\mathcal{L}_{\text{MIPD}_1} + k\,\nabla_{\theta_1}\mathcal{L}_{\text{MIPD}_2}\right), \tag{12}$$

$$\nabla_{\theta_2}\mathcal{L}_{\text{MIPD}} = \lambda_M \left(\nabla_{\theta_2}\mathcal{L}_{\text{MIPD}_2} + k\,\nabla_{\theta_2}\mathcal{L}_{\text{MIPD}_1}\right).^{3} \tag{13}$$

where λ_M is a Lagrange multiplier, and $k \in \{-1, 0, 1\}$ sets the modality's strategy:

- **Collaborative ($k = 1$):** All modalities work together to increase each other's contributions. The $\mathcal{L}_{\text{MIPD}}$ terms are applied across all parameters, resulting in $\min_{\theta} \mathcal{L}_{\text{MIPD}}$.
- **Independent ($k = 0$):** Each modality focuses on maximizing its own contribution by optimizing solely its respective $\mathcal{L}_{\text{MIPD}}$ term, leading to $\min_{\theta_i} \mathcal{L}_{\text{MIPD}_i}$.
- **Greedy ($k = -1$):** Each modality seeks to maximize its own contribution by: 1) minimizing its own $\mathcal{L}_{\text{MIPD}}$ term, and 2) maximizing the $\mathcal{L}_{\text{MIPD}}$ terms of other modalities, resulting in a min-max game, $\min_{\theta_i} \max_{\theta_{\neg i}} \mathcal{L}_{\text{MIPD}_i}$[4].

Following the results in Appendix A.8, we adopt the greedy strategy as default, as it showed the most consistent performance in our setting.

3.3 Perturbations

To assess the importance of modality X_1, we define $\mathcal{L}_{\text{MIPD}_1}$, which captures changes in the model's output when X_1 is perturbed (i.e., $\{\tilde{X}_1, X_2\}$ vs. $\{X_1, X_2\}$). Instead of traditional input-space perturbations, which can be computationally expensive and task-dependent, we apply a within-batch permutation $\sigma_e \sim \text{Uniform}(\mathcal{P})$ in the latent space, yielding $\tilde{X}_1 = \sigma_e(X_1)$. This approach avoids extra forward passes and reduces computational and memory overhead. Further analysis of this technique and comparisons with prior methods are provided in Appendix A.10 and A.14.

The complete algorithm is presented in Algorithm 1, with an extension of $\mathcal{L}_{\text{MCR}}$ to M modalities described in Appendix A.7. In Appendix A.9, we analyze various combinations of loss components, revealing that penalizing task-irrelevant information benefits models with extensive SSL pretraining but proves detrimental for those without it.

4 Experiments

4.1 Synthetic Dataset

We create a scenario where mutual information varies, showcasing modality competition. While various factors can contribute to such a phenomenon, we focus on modality informativeness imbalance to motivate our approach.

Data: We generate task-irrelevant information for each modality by sampling $N_1, N_2 \sim \mathcal{N}(0, \mathbf{I})$ and the 5-class label Y_t from a uniform distribution $Y_t \sim \text{Uniform}(5)$. Each modality is converted into a high-dimensional vector using fixed transformations, similar to Liang et al.[32]. We relate both modalities to the label through a linear relationship: $X_1 = N_1 + Y_t$ and $X_2 = N_2 + Y_t$. Data points are distributed in such a way that either both modalities contain label information (Shared Information) or only one of the modalities (Unique Information). In cases where only one modality contains label information, the other modality is defined as $X_1 = N_1$ and $X_2 = N_2$ respectively. We

[3]The parameter set under the loss indicates where backpropagation applies.

[4]The notation $\neg i$ refers to the rest of the modalities except i.

Algorithm 1 Multimodal Training with MCR

Input: Training dataset D with modalities $X_1, X_2, \ldots, X_M$, labels Y_t, multimodal model $f \in \mathcal{F}$, initialized unimodal encoders θ_i, reconstruction model h, $\lambda_{\text{uni}}, \lambda_{\text{M}}$ Lagrangian coefficients:

for each batch $(X_1, .., X_M, Y_t)$ of each epoch **do**
  Compute $\mathcal{L}_{task}(f(X_1, .., X_M), Y_t)$ and $\mathcal{L}_{task}^{uni} = \lambda_{\text{uni}} \sum_{m=1}^{M} \mathcal{L}_{task}(f_m^u(X_m), Y_t)$
  Extract $(Z_1, .., Z_M)$ from $f(X_1, .., X_M)$
  Assess the $\mathcal{L}_{\text{Con}}$ with Eq. 9 and $\mathcal{L}_{\text{CEB}}$ with Eq. 11
  Sample σ_e permutations and compute permuted pairs on the latent space Z
  Pass each pair through the fusion model f_c to get predictions $\tilde{Y}_m$ with modality m permuted
  Compute $\mathcal{L}_{\text{MIPD}}$ using Eq. 8, 13 and selecting k by strategy (default: Greedy)
  Determine $\mathcal{L}_{\text{MCR}}$ from Eq. 5
  Update the model parameters based on
  $\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{uni} + \mathcal{L}_{\text{MCR}}$
end for

vary the percentage of data points with shared and unique information to analyze model performance under different conditions.

Results: Figure 1 shows the performance on synthetic data, comparing our method (MCR) with several baselines. As the shared information S among the modalities decreases, and the unique information of one modality U_1 increases while the U_2 remains constant, we observe a performance drop for all methods. MCR maintains the highest accuracy across all combinations, demonstrating the slowest decline and highlighting its robustness to such imbalance.

4.2 Real-World Datasets

Datasets: We explore several real-world datasets, primarily with video, optical flow, audio, and text modalities, that either exhibit significant imbalance among modalities or serve as standard multimodal benchmarks. Detailed descriptions are provided in Appendix A.3, with brief summaries below:

1. CREMA-D [4]: An emotion recognition dataset with 91 actors expressing 6 distinct emotions.
2. AVE [51]: A collection of videos with temporally aligned audio-visual events across 28 categories.
3. UCF [47]: An action recognition dataset of real-life YouTube videos.
4. CMU-MOSEI [67]: Multimodal sentiment analysis dataset with 23k monologue clips.
5. CMU-MOSI [65]: Multimodal sentiment analysis dataset with over 2k YouTube video clips.
6. Something-Something (V2) [14]: 220k clips of individuals performing 174 hand actions.

Models: We employ a variety of models and backbone encoders to examine the behavior of both smaller-scale models trained from scratch and larger, more complex models pretrained with self-supervised learning (SSL). This combination demonstrates that our method is effective in both limited data scenarios without pretraining and in cases with ample data where the goal is fine-tuning. We utilize ResNet-18 and small-scale Transformers (from thousand to 20M parameters) alongside state-of-the-art models such as Swin-TF [35] and Conformer [13], incorporating backbone encoders like Wav2Vec2, HuBERT, and ViViT, resulting in model sizes approaching 200M parameters. Detailed model configurations for each dataset are provided in Appendix A.4 and experimental details in Appendix A.5. These choices aim to bridge the gap between theoretical work and practical application.

Results: Table 1 reports accuracy comparisons across baseline methods on our evaluation datasets. We highlight two key observations:

a) Most prior methods, including recent multimodal approaches, fail to outperform simpler alternatives such as Ensembles or unimodal encoders (either frozen or finetuned). While some of these methods partially address modality competition, they often fall short of effectively leveraging multimodal data.

b) MCR is the only method that consistently surpasses all baselines across datasets, model architectures, number of modalities, and task types (classification and regression), with an exception on AVE with Conformer model. This consistent advantage highlights MCR's ability to balance

Table 1: Performance comparison of MCR against prior multimodal training methods across six datasets. MCR consistently achieves top results across various modality combinations (V-A, V-T, V-A-T, V-OF) and model architectures. MOSI and MOSEI use three modalities and are trained as regression tasks (converted to binary accuracy); the rest are classification tasks. All baselines were rerun under our evaluation protocol to ensure fair comparison and address data leakage issues identified in previous evaluation setups.

	CREMA-D		AVE		UCF	MOSI		MOSEI		Sth-Sth
	ResNet	Conformer	ResNet	Conformer	ResNet	Transformer				Swin-TF
Method	V-A	V-A	V-A	V-A	V-A	V-T	V-A-T	V-T	V-A-T	V-OF
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: $78.9_{\pm 1.7}$		V: $61.4_{\pm 0.2}$ OF: $50.8_{\pm 0.1}$
Ensemble	$71.7_{\pm 2.2}$	$84.6_{\pm 1.0}$	$70.5_{\pm 0.2}$	$88.4_{\pm 2.2}$	$52.8_{\pm 0.5}$	$70.5_{\pm 2.1}$	$67.2_{\pm 1.5}$	$78.4_{\pm 0.7}$	$77.2_{\pm 0.6}$	$64.6_{\pm 0.2}$
Joint Training	$62.6_{\pm 5.8}$	$74.6_{\pm 2.2}$	$66.7_{\pm 1.5}$	$82.2_{\pm 0.8}$	$47.7_{\pm 1.5}$	$73.0_{\pm 1.3}$	$73.6_{\pm 1.3}$	$80.5_{\pm 0.2}$	$80.8_{\pm 0.3}$	$57.5_{\pm 0.1}$
Multi-Loss	$69.2_{\pm 1.8}$	$82.6_{\pm 0.9}$	$70.1_{\pm 0.9}$	$86.3_{\pm 1.1}$	$51.1_{\pm 1.8}$	$72.1_{\pm 0.4}$	$73.6_{\pm 2.9}$	$80.0_{\pm 0.7}$	$80.2_{\pm 0.5}$	$61.5_{\pm 0.1}$
Uni-Pre Frozen	$72.4_{\pm 1.8}$	$85.0_{\pm 1.8}$	$72.2_{\pm 0.3}$	$87.2_{\pm 2.4}$	$53.0_{\pm 0.9}$	$73.3_{\pm 1.8}$	$72.7_{\pm 1.6}$	$79.9_{\pm 0.5}$	$79.8_{\pm 0.3}$	$64.0_{\pm 0.2}$
Uni-Pre Finetuned	$73.3_{\pm 1.8}$	$82.4_{\pm 2.0}$	$72.5_{\pm 1.3}$	$86.5_{\pm 0.8}$	$53.5_{\pm 1.3}$	$73.1_{\pm 2.3}$	$73.7_{\pm 0.7}$	$80.3_{\pm 0.4}$	$80.3_{\pm 0.2}$	$62.1_{\pm 0.2}$
MSLR [64]	$56.5_{\pm 2.4}$	$77.1_{\pm 2.4}$	$67.3_{\pm 2.2}$	$81.0_{\pm 1.4}$	$50.9_{\pm 3.9}$	×	×	×	×	–
MMCosine [63]	$59.3_{\pm 1.5}$	$74.0_{\pm 0.3}$	$65.0_{\pm 1.4}$	$83.8_{\pm 0.8}$	$47.3_{\pm 4.1}$	×	×	×	×	–
OGM [42]	$65.6_{\pm 3.8}$	$82.4_{\pm 1.0}$	$67.3_{\pm 0.6}$	$79.7_{\pm 1.4}$	$51.8_{\pm 1.9}$	$73.9_{\pm 1.1}$	⊗	$79.7_{\pm 0.6}$	⊗	$57.8_{\pm 0.5}$
AGM [29]	$69.3_{\pm 1.4}$	$78.5_{\pm 1.6}$	$68.4_{\pm 1.1}$	$85.3_{\pm 0.5}$	$51.0_{\pm 1.6}$	$74.0_{\pm 1.3}$	$73.9_{\pm 1.9}$	$79.3_{\pm 0.4}$	$80.2_{\pm 0.3}$	$56.6_{\pm 0.4}$
MLB [27]	$71.9_{\pm 2.2}$	$85.2_{\pm 0.9}$	$71.6_{\pm 0.2}$	$86.7_{\pm 0.3}$	$52.2_{\pm 1.7}$	$72.4_{\pm 1.7}$	$74.2_{\pm 1.7}$	$80.1_{\pm 0.5}$	$80.5_{\pm 0.4}$	$61.6_{\pm 0.2}$
ReconBoost [19]	$69.0_{\pm 2.4}$	$84.8_{\pm 1.8}$	$68.4_{\pm 1.7}$	$86.1_{\pm 0.7}$	$50.2_{\pm 4.0}$	×	×	×	×	$56.1_{\pm 0.3}$
MMPareto [57]	$69.0_{\pm 2.5}$	$83.8_{\pm 0.8}$	$73.0_{\pm 1.3}$	$87.4_{\pm 1.3}$	$51.4_{\pm 2.2}$	$73.4_{\pm 1.0}$	$73.7_{\pm 0.6}$	$79.3_{\pm 0.6}$	$79.5_{\pm 0.8}$	$59.2_{\pm 0.5}$
D&R [59]	$70.6_{\pm 1.3}$	$85.0_{\pm 0.4}$	$72.3_{\pm 1.5}$	**91.0**$_{\pm 0.7}$	$49.3_{\pm 1.0}$	×	×	×	×	$61.7_{\pm 0.2}$
MCR	**76.1**$_{\pm 1.6}$	**85.7**$_{\pm 0.2}$	**73.4**$_{\pm 0.0}$	$88.8_{\pm 1.0}$	**55.2**$_{\pm 1.8}$	**75.2**$_{\pm 1.7}$	**76.5**$_{\pm 1.4}$	**80.8**$_{\pm 0.4}$	**81.1**$_{\pm 0.4}$	**65.0**$_{\pm 0.1}$

× method not applicable to regression tasks;
⊗ method not applicable to trimodal inputs;
– result not reported.

modality contributions during training under diverse settings, positioning it as a strong and generalizable approach for multimodal learning.

4.3 Analysis of Multimodal Error

Methodology: To understand how MCR improves over existing methods, we perform a post-hoc error analysis by categorizing each sample based on the correctness of the unimodal predictions. Specifically, we consider four groups: (1) both unimodal models are correct, (2) only the first is correct, (3) only the second is correct, and (4) both are incorrect. This breakdown allows us to compare how different multimodal methods behave across these categories and identify whether gains arise from selective reliance on unimodal cues or from synergistic integration.

Results: The error analysis of Figure 4 reveals key strengths and limitations of the proposed method. MCR consistently outperforms other methods in routing information favouring both modalities in the cases that only one of the modalities is correct maintaining competitive performance on both of them and in the case that all modalities correctly predict the label. This demonstrates MCR's ability to effectively route information to the appropriate modality, in line with its design choice to model and control training via this modality-independent, task-relevant information through the mutual information terms.

However, MCR does not exhibit significant gains in capturing synergetic information in the datapoints that all unimodal models fail, underperforming relative to AGM and MLB. This suggests MCR excels in routing decisions but may be less effective at leveraging synergies across modalities when all individual models falter. This trend holds across other datasets (see Appendix A.11), with the exception of MOSI, where MCR improves synergy. Our initial assumption that concurrent modality training would foster synergy thus did not hold in practice.

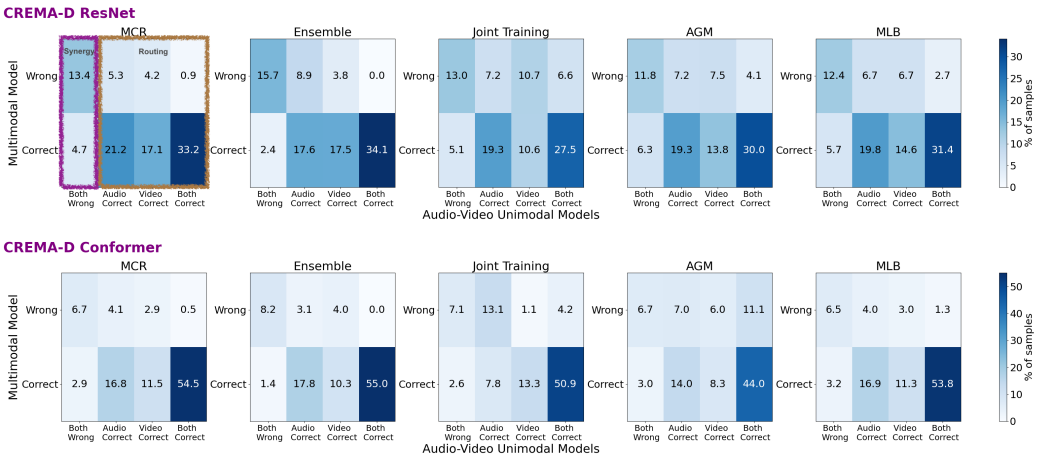


Figure 4: Error comparison on the CREMA-D dataset across unimodal and multimodal models (MCR, Ensemble, Joint Training, AGM, MLB). Each matrix summarizes model performance based on unimodal prediction correctness. MCR performs best when at least one unimodal branch is correct (brown box), effectively preserving modality-specific signals. However, AGM and MLB outperform MCR when both unimodal predictions fail, in the "Both Wrong" (purple box), indicating stronger synergy in those edge cases. Trends across other datasets are shown in Appendix A.11, with MOSI being a notable exception where MCR also excels in synergy.

5 Discussion

This paper examines the challenge of modality competition in multimodal learning, where certain modalities dominate the training process, resulting in suboptimal performance. We introduce the Multimodal Competition Regularizer (MCR), a novel approach inspired by information theory, which frames multimodal learning as a game where each modality competes to maximize its contribution to the final output. MCR efficiently computes lower and upper bounds to optimize both unique and shared task-relevant information for each modality. Our extensive experiments show that MCR consistently outperforms existing methods and simple baselines on both synthetic and real-world datasets, providing a more balanced and effective multimodal learning framework. MCR paves the way for fulfilling the long-standing promise of multimodal fusion methods to achieve performance that surpasses the combined results of unimodal training.

We explored different game strategies and observed that directly encouraging competition between modalities in the overall objective function positively impacts performance, as detailed in Appendix A.8. Future work could investigate more refined strategies to enable individualized and adaptive decisions for each modality to unlock greater performance gains.

Lastly, we conduct a post-hoc error analysis found both in Section 4.3 and Appendix A.11, examining overlaps between the errors of multimodal models and their unimodal counterparts. The results show that MCR excels at routing decisions to the correct unimodal information but does not promote synergetic behavior accordingly, compared to previous methods. Our initial assumption that the simultaneous progress of unimodal encoders during training would naturally enhance synergy was not supported in practice, highlighting the need for future work to promote this behavior explicitly. Finally, this analysis highlights the potential for performance improvements through enhanced multimodal training, motivating further exploration in this area.

Acknowledgement

This material is based upon work partially supported by the FWO Research Project "Task- and device-agnostic Artificial Intelligence (AI) for EEG analysis" (G046925N); C2 "Dissecting agitation in dementia by multimodal sensing (DADS)" (C2M/23/053); HORIZON-HLTH-2022-STAYHLTH "Artificial intelligence-based Parkinson's disease risk assessment and prognosis (AI-PROGNOSIS)" under Grant Agreement No. 101080581; and the Flemish Government (AI Research Program).

References

[1] Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301. PMLR, 2019.

[2] 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.

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

[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] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023.

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

[8] Ian Fischer and Alexander A Alemi. Ceb improves model robustness. *Entropy*, 22(10):1081, 2020.

[9] Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-specific learning rate control for multimodal classification. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pages 412–422. Springer, 2020.

[10] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33:3197–3208, 2020.

[11] 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.

[12] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso. Versatile audiovisual learning for handling single and multi modalities in emotion regression and classification tasks. *ArXiv e-prints (arXiv:2305.07216)*, pages 1–14, May 2023. doi: 10.48550/arXiv.2305.07216.

[13] 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.

[14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[15] 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.

[16] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[17] 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.

[18] 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.

[19] 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*, ICML'24. JMLR.org, 2024.

[20] 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.

[21] Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu, and Si Shen. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6755–6764, 2022.

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[23] Emmanouil Kariotakis, Nikolaos Sidiropoulos, and Aritra Konar. Fairness-regulated dense subgraph discovery. *arXiv preprint arXiv:2412.02604*, 2024.

[24] 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.

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[28] Konstantinos Kontras, Christos Chatzichristos, Huy Phan, Johan Suykens, and Maarten De Vos. Core-sleep: A multimodal fusion framework for time series robust to imperfect modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[29] 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.

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

[31] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

[32] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[37] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[38] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[39] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[40] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[42] 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.

[43] Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

[44] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. *arXiv preprint arXiv:2111.01936*, 2021.

[45] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5213–5224, 2023.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[47] 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.

[48] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

[51] 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 (ECCV)*, pages 247–263, 2018.

[52] 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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

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

[54] 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 (ECCV) Workshops*, pages 0–0, 2018.

[55] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017.

[56] 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.

[57] Yake Wei and Di Hu. Mmpareto: boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[58] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27338–27347, 2024.

[59] 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.

[60] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

[61] 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.

[62] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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 (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[68] 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.

[69] 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: We believe the abstract and introduction accurately reflect the ideas, contribution and the scope of our presented research.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have conducted a series of experiments in the error analysis section that provide the limitations of our method, which are also included in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: [Yes]

Justification: We believe we provide the full set of assumptions and the complete and correct proofs for our methodology in the sections that this applies.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: We have included all the experimental details needed to reproduce our results and claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: In the Appendix we provide a full description with the instructions needed to reproduce our experimental results, from the experimental details, dataset preparation and time needed. We believe these will be sufficient but after the review process we will make our code publicly available too.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have included all the necessary details in the Appendix to make sure someone can follow the training and test details. Since we faced issues with previous works, we believe this is an important part of our contribution.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

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: The paper report error bars in terms of standard error (STE) for the different folds which involve cross validation in the datasets without fixed test set and multiple runs with different seeds in the case of fixed test set. We do not report ste in two cases, in Something-Something dataset, due to its size running multiple folds was computationally prohibitive, and on some instances that we tried due to the size of the test set, the STE was also negligible, the latter reason lead us to skip the reporting of deviation on MOSEI and MOSI. To the rest of our results, we always include STE reporting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: In the corresponding implementation details section we do report about the maximum computational resources used and the execution time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 we align completely with the Code of Ethics of NeurIPS.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

Answer: [NA]

Justification: We believe our work is foundational research which is not tied to a specific task or use and therefore there are no direct societal impacts of the worked performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We believe that our paper does not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: We have properly cited and acknowledged the data, code and models used in our work.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: [NA]

    Justification: Our work does not involve crowdsourcing or research with with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: [NA]

Justification: Our work does not involve crowdsourcing or research with with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our work does not include LLM as core method for developing our research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Supplementary Material

A.1 Evidence of limitations of supervised multimodal training

Identifying the instances where supervised multimodal training collapses is often easier than resolving the issue itself. Nevertheless, it is crucial to understand both these limitations and why simple solutions might not suffice. To explore this, we utilize a ResNet-18 [17] backbone on the CREMA-D dataset, employing audio and video as the two modalities. These modalities are concatenated just before the final linear layer. We measure multimodal performance at the end of each epoch and, simultaneously, perform linear probing on each modality to evaluate their individual contributions throughout training.

In Figure 5, it is clear that the performance of the multimodal model aligns closely with that of the audio modality alone, suggesting that the model heavily relies on audio while neglecting the video modality. This lack of exploration results in the video modality remaining at chance-level accuracy throughout training. As a result, the model fails to leverage any information available in the video modality and performs significantly worse than an ensemble of the unimodally trained models.

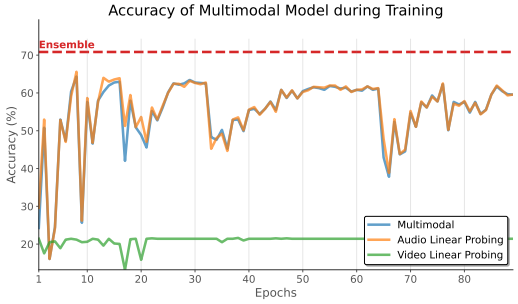


Figure 5: Accuracy of the multimodal model on the CREMA-D dataset across training epochs, showing the performance of the full multimodal model (blue) and individual modality linear probing for audio (orange) and video (green). The dashed red line represents the accuracy of a unimodal ensemble model, highlighting how the model's over-reliance on the audio modality negatively impacts the utilization of the video modality

A.2 Multimodal Competition Error

Multimodal competition occurs when the network primarily optimizes for one modality, leading to a decline in generalization. One modality dominates the reduction in training error and limits gradient feedback to the other modalities. Huang et al. [20] proved that in late fusion models (i.e. $\theta_c = \emptyset$), each modality has a probability of causing multimodal competition. They define the phenomenon in terms of the correlation $\Gamma_m = \sum_{cl} \max[\langle X_m^{cl}, W^{cl} \rangle]^+$ of each input modality and the weights, W, of the output layer for each corresponding class cl. If a modality dominates the competition, it causes the other modality to maintain its initial (i.e. before training) correlation levels. Such correlation levels do not have a specific target, unlike other problems where equality is the goal [23], making the solution less straightforward. We proceed with this analysis on late-fusion models by introducing the generalization error ϵ resulting from the effects of multimodal competition.

Definition A.1. Definition of Multimodal Competition Error (MCE): Let Γ_1 and Γ_2 represent the correlations of modalities Z_1 and Z_2 with the output. If, after training, $\Gamma_1 \gg \Gamma_2$, and modality X_2 has predictive power greater than randomness ($R(X_2) < R(\text{random})$), while making errors independent from those of X_1 ($\mathbb{E}[e_1 \cdot e_2] = 0$, where e_1 and e_2 are the errors of X_1 and X_2, respectively), then for a trained model f and its optimal solution f^*, there exists a multimodal competition error ϵ, such that:

$$\epsilon \geq R(f^*) - R_{\text{emp}}(f^*) \text{ and } \epsilon \leq R(f) - R_{\text{emp}}(f) \tag{14}$$

where $R(\cdot) = \mathbb{E}[\mathcal{L}_{\text{task}}(Y, \cdot)]$ and $R_{\text{emp}}(\cdot)$ denote the generalization and empirical model risks, and $\mathcal{L}_{\text{task}}$ represents the corresponding task loss, which could vary depending on the objective, such as cross-entropy for classification or mean-squared error for regression. A higher ϵ indicates a stronger effect of multimodal competition, implying that the dominance of one modality significantly impacts the model's generalization. These inequalities are empirically observed through improvements in

generalization achieved by adjusting the training objectives to address multimodal competition, without altering the model or the data. Numerically estimating ϵ would require knowledge of the optimal solution f^*, which is typically unavailable. Lastly, transitioning from the late-fusion models of Huang et al. [20] we need to extend this definition by replacing correlation with MI to consider also non-linear statistical dependencies.

A.3 Datasets

CREMA-D [4]: is an emotion recognition dataset with audio and video modalities. It features a diverse group of 91 actors, covering a wide range of ages, ethnicities, and genders. To ensure consistency, each actor is positioned at an equal distance from the camera, expressing six distinct emotions: Happy, Sad, Anger, Fear, Disgust, and Neutral. In alignment with methodologies from prior studies [42, 29, 6], video frames are sampled at 1 fps, selecting 3 consecutive frames, while audio segments are sampled at 22 kHz, capturing 3 seconds that correspond with the chosen video frames. Audio analysis utilizes a window size of 512 and a step size of 353 samples for Short-Time Fourier Transform (STFT), creating log-Mel spectrograms. For advanced models, our methodology aligns with Goncalves et al. [12], incorporating audio signals sampled at 16 kHz and utilizing pre-calculated facial features. Unlike previous approaches [42, 29, 6], our dataset division follows Goncalves et al. [12], excluding actor overlap between training, validation, and test sets. We report standard deviation (std) across folds for consistency.

AVE [51]: contains 4143 videos across 28 event categories with a wide range such as frying food or playing guitar, each with temporally labeled audio-visual events of at least 2 seconds. Following [6], video segments where the event occurs are sampled at 1 fps for 4 frames, with audio resampled at 16 kHz using CREMA-D's STFT settings. AVE provides predefined training, validation, and test splits. Our std is derived from three random seeds on the same test set.

UCF101 [47]: features real-life action videos from YouTube in 101 action categories, expanding on UCF50. We select samples that include both video and audio modalities, narrowing our focus to 51 action categories. Data preparation mirrors that of the AVE dataset. For model evaluation, we utilize the 3-fold split offered [47], reporting the std across these folds.

Something-Something (V2) [14]: presents 220,847 video clips where individuals execute 174 distinct, object-agnostic hand actions, spanning a wide range of simple hand movements without reliance on specific objects. This extensive collection markedly exceeds the data volume of prior datasets. In alignment with Radevski et al. [45], we integrate Optical Flow (OF) as the additional modality for capturing dynamic motions. We sample videos at 1 fps, retaining 16 frames per sample. Our data preparation adheres to the pipeline outlined in Radevski et al. [45]. Results are reported based on three random seeds on the same validation set.

CMU-MOSI [65] and -MOSEI [67]: datasets serve as benchmarks for multimodal sentiment and emotion analysis. MOSI comprises 2,199 video clips with video, audio and text modalities. MOSEI expands on this with around 10x more YouTube movie review clips from 1000 different speakers. Both datasets are annotated with sentiment scores (-3 to 3) and following previous works [52] we employ metrics such as 7-class accuracy, binary accuracy, F1 score, mean absolute error, and correlation with human annotations for model evaluation. We use for both datasets the aligned versions following [31]. The reported std is derived from three random seeds on the same test set.

A.4 Models: Backbone Unimodal Encoders

In line with previous research [29, 42, 6, 63], our initial experiments adopt ResNet-18 [17] as the unimodal encoder for handling both video and audio modalities in the CREMA-D, AVE, and UCF datasets. These models are randomly initialized and incorporate adaptive pooling to accommodate diverse input dimensions. For the CMU-MOSEI dataset, we exploit a 5-layer Transformer [53] similar to previous works [32, 31].

We extend our investigation to include a larger, pre-trained set of unimodal encoders that are optimally suited for each specific modality. This selection process is designed to rigorously assess whether state-of-the-art models exhibit susceptibility to the same phenomena under investigation. We exploit these encoders on the CREMA-D dataset. Following [13] on CREMA-D, we deploy the first 12 layers of the Wav2Vec2 [3, 60] model with self-supervised pretrained weights for speech recognition,

allowing the Wav2Vec2 model to be finetuned. For the video modality, we extract the facing bounding boxes with the multi-task cascaded convolutional neural network (MTCNN) face detection algorithm [68] and afterward the facial features of every available frame exploiting EfficientNet-B2 [50] as a frozen feature descriptor. The extracted audio features and pre-calculated facial features are further refined using a 5-layer Conformer [15], initialized from scratch. The total model size is 183M parameters. Although the model includes additional components, we refer to it as "Conformer" for simplicity.

For the AVE dataset, we use a similar architecture to CREMA-D, where each branch utilizes an advanced, pretrained model aligned with AVE data, followed by a 5-layer Conformer. For the video branch, we use the ViViT model [2] pretrained on Kinetics [24], and for the audio branch, the HuBERT [18] model pretrained on Audioset [11] resulting in a model with 232M parameters. Ensuring the audio pretraining included non-speech data was important for the pretraining to be beneficial. We also refer to this model as "Conformer," although it incorporates different large pretrained models in this instance.

In the case of the Something-Something dataset, our methodology builds upon the insights presented by Radevski et al. [45], which highlight the importance of modality-specific processing in multimodal tasks. For both video and optical flow data, we adopt the Swin Transformer [35] as the backbone encoder to each modality. This state-of-the-art architecture excels in capturing hierarchical and spatiotemporal features through its shifted window attention mechanism.

A.5 Experimental Details

In this section, we outline the necessary details to reproduce our experiments. Across all datasets, we follow a consistent procedure: we first determine an appropriate learning rate (lr) for the unimodal models by testing several candidates until finding one that works across both modalities. While this step could be avoided by exploiting parameter-specific learning rates, we expected stability implications which we aimed to avoid. For all the experiments of the same dataset/model pair, we use the same hyperparameters, except when fine-tuning pretrained encoders, where we apply a learning rate scaled one magnitude lower.

All models are optimized using Adam [26] with a cosine learning rate scheduler and a steady warm-up phase, except for the Something-Something dataset, where we use Adaw [36]. Early stopping is applied for all models, with maximum epochs set to 100 for ResNet and Transformer models, 50 for Conformer models, and 30 epochs in total for Swin Transformers without early stopping. Batch sizes are adjusted based on computational resources, with ResNets and Transformers both using a batch size of 32, Conformers using 8, and Swin-TF using 16. These settings ensure balanced performance and efficient training across all experiments. In several instances, initializing the encoders with pre-trained weights from unimodal training proved beneficial. This was particularly effective for datasets without precomputed features and models without access to larger-scale SSL pretraining. We use different learning rates (lr) and weight decay (wd) values across experiments, tailored to each dataset and model. For ResNet models, we use $lr = 1\text{e}{-3}$ and $wd = 1\text{e}{-4}$ for CREMA-D, while both AVE and UCF use $lr = 1\text{e}{-4}$ and $wd = 1\text{e}{-4}$. For Transformer models, including MOSI and MOSEI on two and three modalities, the hyperparameters are consistent $lr = 1\text{e}{-4}$ and $wd = 1\text{e}{-4}$. Similarly, for Conformer models, we set $lr = 5\text{e}{-5}$ and $wd = 5\text{e}{-6}$ for CREMA-D, while AVE uses $lr = 1\text{e}{-4}$ and $wd = 1\text{e}{-4}$. Finally, for Swin-TF models trained on the Something-Something dataset, we configure $lr = 1\text{e}{-4}$ and $wd = 0.02$.

Each of the previous methods includes its own set of hyperparameters, typically just one, with some exceptions such as MSLR or D&R, which requires additional parameters. For each dataset/model combination, we conduct a brief hyperparameter search, ensuring an equitable number of trials across methods. Due to the extensive list of hyperparameters, we will provide detailed configurations for each experiment in our GitHub repository. The repository link will be included here following the double-blind review process.

Lastly, all of our experiments run on single GPU with different nodes being used for different experiments. For the largest ones we utilized H100 with 80Gb vram to run the experiments of Sth-Sth which required the longest of all up to 48 hours per run. For the rest, we would have from some minutes on the smallest experiment up to 4-5 hours for the datasets CREMA-D, AVE and UCF depending on the GPU and the available RAM.

A.6 Bounding task-irrelevant information via CEB

Our objective is to maximize the information shared between modalities that is irrelevant to the supervised task. Directly estimating the conditional mutual information $I(X_1; X_2 \mid Y)$ is challenging in high-dimensional settings, so we relax the objective by penalizing both the individual and shared irrelevant information, leading to the decomposition:

$$I(X_1; X_2 \mid Y) = H(X_1 \mid Y) + H(X_2 \mid Y) - H(X_1, X_2 \mid Y) \tag{15}$$

$$\Rightarrow -I(X_1; X_2 \mid Y) + H(X_1 \mid Y) + H(X_2 \mid Y) = H(X_1, X_2 \mid Y) \tag{16}$$

We lower-bound the conditional joint entropy using CEB [7] as follows,

$$H(X_1, X_2 \mid Y) = -\mathbb{E}_{p(x_1,x_2,y)} \left[\log p(x_1, x_2 \mid y)\right] \tag{17}$$

$$= -\mathbb{E}_{p(x_1,x_2,y)} \left[\log g(x_1, x_2 \mid y)\right] - \mathrm{KL}\big(p(x_1, x_2 \mid y) \,\|\, g(x_1, x_2 \mid y)\big) \tag{18}$$

$$\leq -\mathbb{E}_{p(x_1,x_2,y)} \left[\log g(x_1, x_2 \mid y)\right] \tag{19}$$

Assuming a conditional Gaussian model $g(x_1, x_2 \mid y) = \mathcal{N}((x_1, x_2); \mu(y), \sigma^2 I)$, we define $\mu(y)$ as a deterministic function $h : Y; \theta_h \to Z = (Z_1, Z_2)$ that predicts a target joint representation from Y. Then, the conditional entropy term is upper-bounded as:

$$H(X_1, X_2 \mid Y) \leq -\mathbb{E}_{p(x_1,x_2,y)} \left[\log g(x_1, x_2 \mid y)\right] \propto \mathbb{E}_{p(x_1,x_2,y)} \left\|[f(x_1), f(x_2)] - h(y; \theta_h)\right\|^2 \tag{20}$$

Thus, we approximate the entropy term with an MSE loss, encouraging (x_1, x_2) to deviate from any deterministic function when it is not informative for the task. Maximizing $H(X_1 \mid Y)$ and $H(X_2 \mid Y)$ incentivizes each modality to retain information that is task-independent, aligning with the goal of minimizing $I(X_1; X_2 \mid Y)$ and suppressing task-irrelevant alignment between modalities.

A.7 MCR on M modalities

In this section we provide the analysis of MCR for M number of modalities. In that case, the total mutual information $I(X_1, \ldots, X_M; Y)$ can be decomposed into contributions from individual modalities and their subsets as:

$$I(X_1, .., X_M; Y) = \sum_{\mathcal{S} \subseteq \{X_1, .., X_M\}, \mathcal{S} \neq \emptyset} I(\mathcal{S}; Y \mid \{X_1, \ldots, X_M\} \setminus \mathcal{S}) + I(X_1, .., X_M) - I(X_1, .., X_M \mid Y), \tag{21}$$

where $\mathcal{S} \subseteq \{X_1, .., X_M\}$ represents a subset of all modalities, excluding the empty set ($\mathcal{S} \neq \emptyset$). The term $\{X_1, .., X_M\} \setminus \mathcal{S}$ denotes the complement of $\mathcal{S}$, capturing the set of modalities not included in $\mathcal{S}$. The mutual information $I(\mathcal{S}; Y \mid \{X_1, .., X_M\} \setminus \mathcal{S})$ quantifies the information shared between the subset $\mathcal{S}$ and the target variable Y, conditioned on the remaining modalities. This formulation ensures that all modalities, along with their combinations, are accounted for in the summation. It comprehensively captures interactions at every granularity, from individual modalities ($|\mathcal{S}| = 1$) to the full set of modalities ($|\mathcal{S}| = M$).

While we experimented with reducing the number of terms by considering only the cases where $|\mathcal{S}| = 1$, we observed a slight improvement in performance when including all terms for three modalities. However, as the number of modalities increases, it might be beneficial to sub-select and exclude certain terms to mitigate the computational burden and prevent an overflow of terms.

A.8 Ablation Study - Game Strategies

We perform an ablation study to test our hypothesis that framing multimodal competition regularization as a game benefits the model by avoiding destructive loss interactions in each backbone encoder. Table 2 compares among the three strategies: Collaborative, Independent, and Greedy. The results show that, across the models allowing backbone encoders to maximize their own CMI term and concurrently minimizing the others (Greedy strategy) consistently yields the best performance. This result demonstrates that framing multimodal models as competing modalities using game-theoretic principles in the loss terms can be beneficial in balancing these loss terms.

Table 2: Ablation Study: Game Strategies – Comparison of model accuracy on CREMA-D dataset for different game strategies: Collaborative, Independent, and Greedy, using both ResNet and Conformer backbones.

Dataset/Model Setting	Collaborative $\min_{\theta} \mathcal{L}_{\text{MIPD}}$	Independent $\min_{\theta_i} \mathcal{L}_{\text{MIPD}_{X_i}}$	Greedy $\min_{\theta_i} \max_{\theta_{\neg i}} \mathcal{L}_{\text{MIPD}_{X_i}}$
CREMA-D ResNet + MCR	73.4±3.0	76.0±2.0	**76.2±1.7**
CREMA-D Conformer + MCR	82.9±0.7	82.6±2.6	**85.7±0.2**
AVE ResNet + MCR	67.9±2.5	72.5±1.0	**73.4±0.0**
AVE Conformer + MCR	**88.9±1.2**	88.6±1.1	88.8±1.0
UCF ResNet + MCR	55.1±0.1	54.8±1.6	**55.2±1.8**
MOSI V-T TF + MCR	73.7±1.3	73.6±1.7	**75.2±1.7**
MOSI V-A-T TF + MCR	75.7±2.1	74.4±1.1	**76.5±1.4**
MOSEI V-T TF + MCR	80.4±0.5	80.4±0.5	**80.8±0.4**
MOSEI V-A-T TF + MCR	80.7±0.2	80.8±0.2	**81.1±0.4**
Sth-Sth SwinTF + MCR	64.9±0.1	64.9±0.1	**65.0±0.1**

A.9 Ablation Study - Loss Components

Table 3 presents the model's performance comparison when different loss components are applied. The models utilize pretrained initialization: ResNet with unimodal pretraining and Conformer with SSL. Two key observations can be made:

1. The concurrent exploitation of both $\mathcal{L}_{\text{MIPD}}$ and $\mathcal{L}_{\text{Con}}$ is yielding consistent improvement. Exploiting them separately leads to smaller improvement for $\mathcal{L}_{\text{Con}}$ and even to a decline for $\mathcal{L}_{\text{MIPD}}$. suggests that alignment in the latent space between the modalities is necessary for the permutations to be effective.
2. The $\mathcal{L}_{\text{CEB}}$ term, which penalizes task-irrelevant information, improves the Conformer model's performance, likely due to its pretraining on large, unlabelled datasets that introduce such irrelevant information. In contrast, for the ResNet model, where pretraining already focuses on task-related information, the $\mathcal{L}_{\text{CEB}}$ term does not provide additional benefits.

Table 3: Ablation Study: Regularization Components – Accuracy (%) of ResNet and Conformer models across datasets with different combinations of MCR components: $\mathcal{L}_{\text{MIPD}}$, $\mathcal{L}_{\text{Con}}$, and $\mathcal{L}_{\text{CEB}}$. Results indicate that combining $\mathcal{L}_{\text{MIPD}}$ and $\mathcal{L}_{\text{Con}}$ is crucial for improvement, while $\mathcal{L}_{\text{CEB}}$ does not benefit all models.

MCR Components			ResNet			Conformer		SwinTF
$\mathcal{L}_{\text{MIPD}}$	$\mathcal{L}_{\text{Con}}$	$\mathcal{L}_{\text{CEB}}$	CREMA-D	AVE	UCF	CREMA-D	AVE	Sth-Sth
			73.4±2.5	71.1±1.4	50.0±2.0	84.1±0.6	87.9±1.1	64.7±0.1
✓			74.1±2.9	72.1±0.5	49.4±1.9	83.9±1.8	87.8±1.7	64.8±0.2
	✓		73.4±2.1	72.6±0.6	54.8±1.2	84.5±0.3	88.7±1.4	64.7±0.1
✓	✓		**76.2±1.7**	**73.3±0.5**	**55.1±0.6**	84.5±0.3	88.7±1.4	64.8±0.1
✓	✓	✓	75.6±1.9	72.1±0.9	54.7±1.1	**85.7±0.2**	**88.8±1.0**	**65.0±0.1**

A.10 Ablation Study - Perturbation methods

To estimate the importance of a modality, we perturb one modality while keeping the other fixed and observe the change in the model's output. Several prior methods have proposed ways to do this, but each comes with trade-offs. Additive noise has been used to maximize output variance as a proxy for functional entropy [10], though this increases sensitivity to noise, conflicting with goals such as smoothness and robustness [49, 1]. Task-specific augmentations [21, 32] rely on handcrafted strategies that may not generalize across domains or modalities. Zero-masking strategies used for approximating Shapley values [29] are theoretically grounded but often unreliable in high-

dimensional settings [37] and require multiple forward passes, increasing computational and memory demands.

Based on these previous works we explore three types of perturbation methods to analyze their impact on the performance of MCR: noisy perturbations, zero-masking, and permutations. Each method was applied in different spaces (input space and latent space) or within the batch structure to determine how effectively MCR can leverage these perturbations to enhance multimodal learning. In Table 4, we summarize the different approaches we examine and in Table 5 we present the results of an ablation study comparing the performance of MCR under these perturbation techniques across multiple datasets: CREMA-D, AVE, UCF, MOSEI, and MOSI.

Table 4: Overview of the approaches examined for the permutation methods.

Noise in the Input Space	Adding noise directly to the input features of each modality, simulating realistic data corruption. For its implementation we follow [10].
Shapley Input-Space Perturbations	Following the approach of [29], Shapley zero-induced values are used to determine the importance of input modalities.
Noise in the Latent Space	Applying noise to the latent representations and encouraging robustness at the feature extraction level.
Zeros in the Latent Space	Zero-masking latent representations to disrupt one modality.
Within-Batch Permutations in the Latent Space	Permuting data points within the batch to disrupt alignment.

Table 5: Ablation study comparing different perturbation methods for MCR across multiple datasets. The table shows the performance of MCRwhen combined with various perturbation techniques, including input or latent space noise, Shapley values in the input space, and within-batch permutations.

Method	CREMA-D	AVE	UCF	MOSEI	MOSI
MCR with Noise Input-Space	75.3$_{\pm 2.9}$	72.1$_{\pm 1.1}$	54.6$_{\pm 0.8}$	80.5$_{\pm 0.4}$	74.7$_{\pm 0.1}$
MCR with Shapley Input-Space	73.6$_{\pm 2.5}$	72.6$_{\pm 0.9}$	**55.5$_{\pm 0.6}$**	79.8$_{\pm 0.5}$	74.3$_{\pm 2.2}$
MCR with Noise Latent-Space	73.6$_{\pm 1.1}$	72.6$_{\pm 0.4}$	54.5$_{\pm 0.7}$	80.1$_{\pm 0.6}$	72.3$_{\pm 1.7}$
MCR with Zero Latent-Space	73.6$_{\pm 1.9}$	**73.3$_{\pm 0.5}$**	54.5$_{\pm 0.4}$	79.6$_{\pm 0.4}$	73.6$_{\pm 2.4}$
MCR with Permutations Latent-Space	**76.1$_{\pm 1.1}$**	**73.3$_{\pm 0.5}$**	55.2$_{\pm 1.8}$	**80.8$_{\pm 0.4}$**	**75.2$_{\pm 1.7}$**

We observe that Shapley-based input-space perturbations show competitive performance, particularly in datasets like UCF, MOSI, and MOSEI, while noise-based methods (both input and latent spaces) achieve reasonable performance, they consistently underperform other techniques. While input-space perturbations could be a viable option, they significantly increase computational complexity, as they require an additional forward pass through the typically large unimodal encoders for each sample. This limitation, which we analyze in Appendix A.14, makes them less favorable as a practical solution. Finally, these findings support the choice of permutations as the preferred perturbation method, while suggesting that further exploration of alternative strategies could potentially lead to even greater improvements.

The semantic meaning of this perturbation depends on whether the permuted sample shares the same label as the original. If the labels match, the perturbation is semantically valid and can be seen as an implicit augmentation. In this case, a large output change indicates that the model may be relying on spurious or unstable features within the modality. If the labels do not match, the resulting input is semantically inconsistent and can be interpreted as out-of-distribution. If the model's output is insensitive to such a perturbation, this suggests that the modality is being ignored. Conversely, a sensitive reaction may indicate an overreliance on features not robust to semantic shifts. Therefore each category of permuted samples contributes differently to the final output. In practice, we apply both semantically consistent and inconsistent permutations during training. This choice introduces minimal computational overhead and appears to slightly improve convergence stability.

We note that the gradients $\nabla_{\theta_1}\mathcal{L}_{\mathrm{MIPD}_2}$ and $\nabla_{\theta_2}\mathcal{L}_{\mathrm{MIPD}_1}$ can negatively impact model robustness depending on the type of perturbation applied. When perturbations yield out-of-distribution unimodal inputs, such as zero-masking or additive noise, the resulting gradients may encourage the model to learn spurious patterns. In contrast, our main experiments use within-batch permutations, which preserve the in-distribution structure of each modality. Under permutation as the perturbation method, the $\mathcal{L}_{\mathrm{MIPD}}$ formulation remains symmetric regardless of which modality is perturbed, and the gradients contribute constructively to learning in all branches.

A.11 Analysis of Multimodal Error

We extend the error analysis from Section 4.3 by comparing unimodal and multimodal predictions in Figure 6. The results echo the pattern seen in CREMA-D (Figure 4): MCR excels when at least one unimodal model predicts correctly, but still trails MLB and AGM when all unimodal models fail. An exception is the MOSI dataset, where MCR performs well in synergy, even with three modalities.

A.12 Statistical Importance

We assess the statistical significance of the performance differences between our method and each baseline using the Wilcoxon Signed-Rank Test, applied to per-dataset average results. To control for multiple comparisons, we apply the Holm correction to the resulting p-values. We consider results significant if the adjusted p-value is below $\alpha = 0.05$. Full results are shown in Table 6.

Table 6: Wilcoxon Signed-Rank Test results (two-sided) comparing our method to each baseline across datasets. Statistically significant comparisons after Holm correction ($\alpha = 0.05$) are bolded.

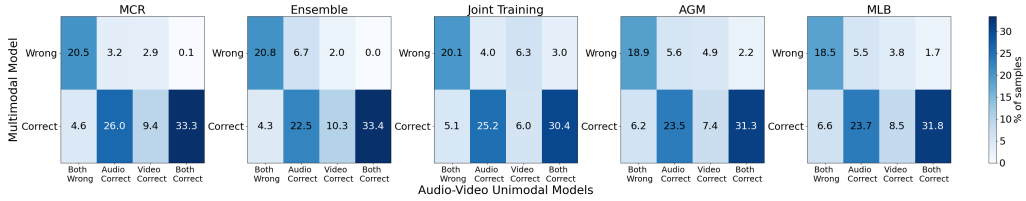
Comparison	**Raw p-value**	**Holm-adjusted p-value**
MCR vs Ensemble	**0.00195**	**0.02148**
MCR vs Joint Training	**0.00195**	**0.01074**
MCR vs Multi-Loss	**0.00195**	**0.00716**
MCR vs Uni-Pre Frozen	**0.00195**	**0.00537**
MCR vs Uni-Pre Finetuned	**0.00195**	**0.00430**
MCR vs OGM	**0.00781**	**0.00955**
MCR vs AGM	**0.00195**	**0.00358**
MCR vs MLB	**0.00195**	**0.00307**
MCR vs ReconBoost	**0.03125**	**0.03438**
MCR vs MMPareto	**0.00195**	**0.00269**
MCR vs D&R	0.15625	0.15625
Greedy vs Collaborative	**0.00977**	**0.00977**
Greedy vs Independent	**0.00195**	**0.00391**
L_{MIPD} + L_{Con} + L_{CEB} vs No Reg	**0.03125**	0.12500
L_{MIPD} + L_{Con} + L_{CEB} vs L_{MIPD}	**0.04311**	0.08623
L_{MIPD} + L_{Con} + L_{CEB} vs L_{Con}	0.06789	0.09052
L_{MIPD} + L_{Con} + L_{CEB} vs L_{MIPD} + L_{Con}	0.84375	0.84375

Dynamics of MIPD Components

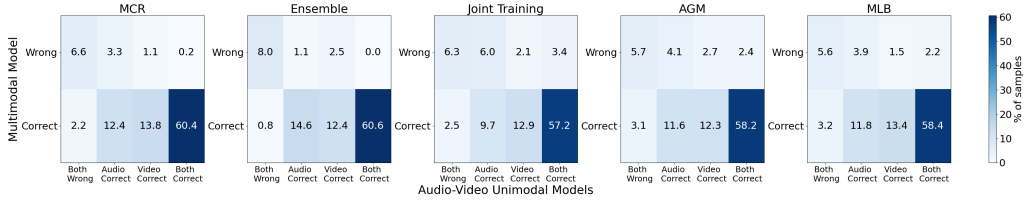
To illustrate the learning dynamics of the MCR regularizer, we can analyze the evolution of its core loss components during training. Figure 7 plots the two MIPD terms, corresponding to the video and text modalities, from a training run on the MOSI V-T dataset.

These MIPD terms, which serve as proxies for the unique contribution of each modality, exhibit a dynamic, alternating behavior. The fluctuations reflect shifts in which modality is more influential on the fused output at different stages of training. This plot visualizes MCR's mechanism for actively balancing modality importance, preventing one from consistently dominating the other. It is important to note that while this alternating pattern is the desired behavior, its specific form and prominence can vary across different training runs and datasets. Therefore, this figure is presented as a clear, illustrative example of the dynamic interplay MCR encourages. This prevents static dominance by a single modality, which is key to mitigating modality competition.

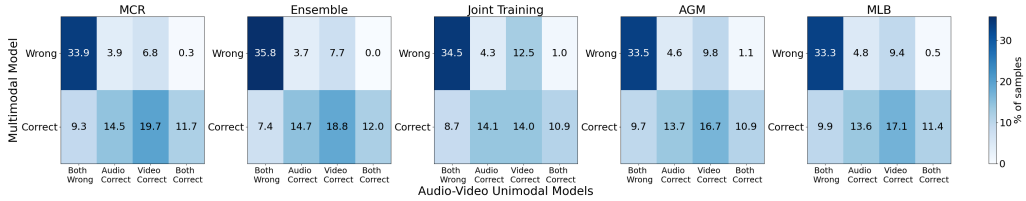
AVE ResNet



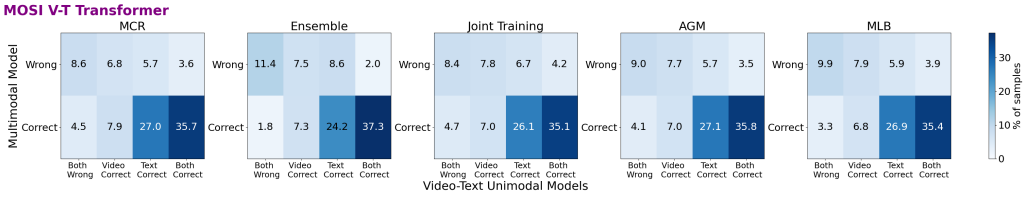
AVE Conformer



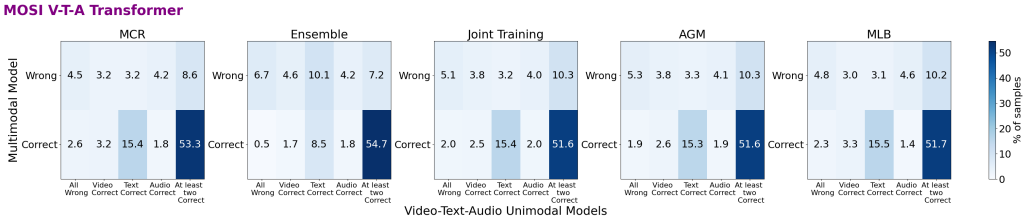
UCF ResNet



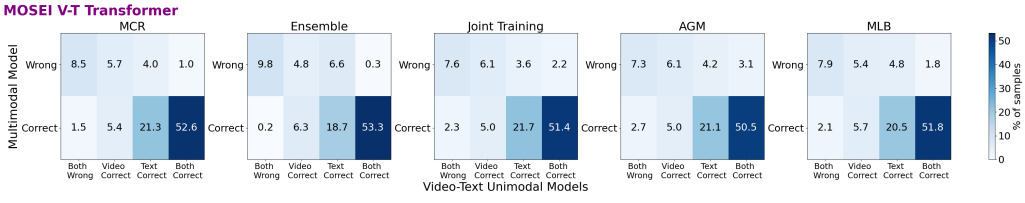
MOSI V-T Transformer



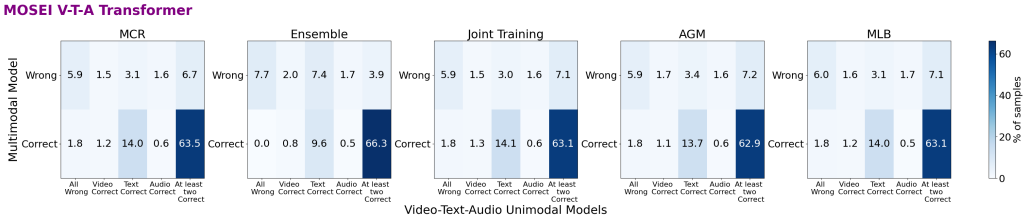
MOSI V-T-A Transformer



MOSEI V-T Transformer



MOSEI V-T-A Transformer



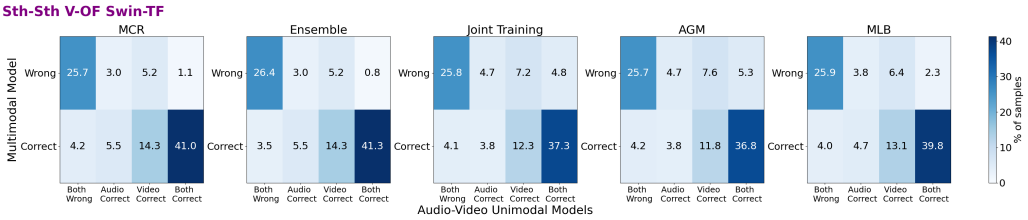


Figure 6: Error comparison matrices across pairs of datasets and models, comparing unimodal predictions with multimodal models trained using various methods, including Ensemble, Joint Training, AGM, MLB, and MCR for the datasets AVE, UCF, MOSI, MOSEI and Something-Something (Sth-Sth). Each column of the confusion matrix represents cases where both unimodal predictions are incorrect, where only one is correct, and where both are correct. The results highlight that MCR consistently performs well in cases where at least one unimodal prediction is correct. Additionally, MLB and AGM in many instances outperform MCR in discovering synergetic information, which refers to the "Both/All Wrong" column, highlighting a current limitation of MCR.

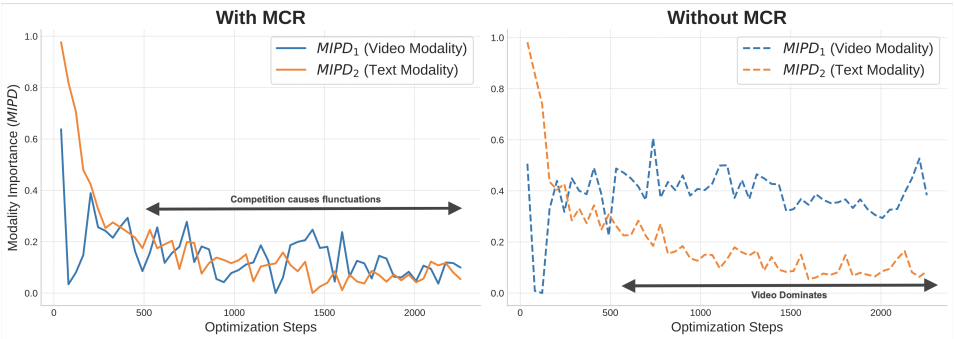


Figure 7: The plots show the two MIPD losses (MIPD_1 and MIPD_2) for each modality, with (left) and without (right) regularization. With regularization, the losses exhibit an alternating pattern, demonstrating MCR's ability to dynamically explore multiple modality contribution combinations and prevent collapse into a single modality. Without regularization, the text modality initially contributes more strongly, but the video modality eventually outperforms and dominates.

A.13 Reproduction Challenges for MLA

We attempted to reproduce the results reported for the MLA method [69] using the authors' officially released code. Despite carefully replicating the training procedure, we were unable to reach close to the reported performance. Specifically, our reproduced accuracy results on our five different settings are as follows: CREMA-D (ResNet) $61.7_{\pm 4.1}$, CREMA-D (Conformer) $75.1_{\pm 3.5}$, AVE (ResNet) $44.5_{\pm 1.9}$, AVE (Conformer) $72.4_{\pm 3.5}$, and UCF (ResNet) $47.9_{\pm 1.4}$. These results are substantially lower than those reported in Table 1 of the original paper. We suspect the discrepancy may be due to missing implementation details. We include these findings in the interest of transparency and to support ongoing efforts toward reproducibility in multimodal learning.

A.14 Computational Speed and Memory Analysis of Perturbations

The computational load imposed by any sample that requires an additional pass can be divided into the encoders f_1, f_2 and the fusion network f_c. The encoders have a computational cost of $cost_{enc}$ per sample, and the fusion network has a cost of $cost_c$ per sample. Thus, the total computational complexity is $\mathcal{O}((M+1) * N * (cost_{enc} + cost_c))$, where M is the times we draw noisy samples to which we add the non-perturbed batch and N is the batch size. If we now use permutation samples that can be directly drawn from the latent space, the additional computational complexity is reduced to $\mathcal{O}(N * cost_{enc} + (M+1) * N * cost_c)$. Compared to the necessity of a supervised forward pass this translates to an addition of $\mathcal{O}(M * N * cost_c)$ computational burden. In most state-of-the-art models, each modality encoder is significantly larger than the fusion network, resulting in $cost_{enc} >> cost_c$.

In such networks, permutations can have almost negligible additional computations. The memory footprint follows a similar pattern.

B Proof of Supervised Contrastive Loss as Lower Bound

We consider the supervised contrastive loss with ψ being the critic function and we rewrite it as follows:

$$\mathcal{L}_{\text{Con}}(X_1, X_2) = \sum_{i \in \mathcal{D}} \frac{-1}{|\mathcal{P}_i|} \sum_{k \in \mathcal{P}_i} \left[\log \frac{\psi(x_{1_i}, x_{2_k})}{\sum_{j \in \mathbb{I}} \psi(x_{1_i}, x_{2_j})} \right] \tag{22}$$

where $\mathcal{P}_i = \{p \in \mathcal{D} \mid y_p = y_i\}$. In supervised contrastive learning, the presence of multiple positive samples turns this into a multi-label problem, unlike traditional noise contrastive estimation (NCE) methods [41], which typically assume only one positive sample. By taking the version of supervised contrastive learning with the expectation over the positives outside of the log, we can interpret each classification as an average of classifiers, with each classifier focusing on identifying one of the positive samples.

For each positive sample $p \sim \mathcal{P}_i$, we aim to derive the optimal probability of correctly identifying that point, denoted $d = p$. This is done by sampling the point from the conditional distribution $p(x_{2_p} \mid x_{1_i}, y_i)$ while sampling the remaining points from the proposal distribution $p(x_{2_l})$. This approach mirrors the technique used in InfoNCE [41] and leads to the following derivation:

$$p(d = p | X_2, x_{1_i}, y_i) = \frac{p(x_{2_p} \mid x_{1_i}, y_i) \prod\limits_{l \in \mathcal{D}, l \neq p} p(x_{2_l})}{\sum_{j \in \mathcal{D}} p(x_{2_j} \mid x_{1_i}, y_i) \prod\limits_{l \in \mathcal{D}, l \neq j} p(x_{2_l})} \tag{23}$$

$$= \frac{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}}{\sum_{j \in \mathcal{D}} \frac{p(x_{2_j}|x_{1_i}, y_i)}{p(x_{2_j})}} \tag{24}$$

The optimal value for the critic function ψ in Equation 24 is proportional to $\psi \propto \frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}$. The MI between the variables can be estimated as follows:

$$L_{\text{Con}}^{Opt} = - \mathop{\mathbb{E}}_{i \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{p \sim \mathcal{P}_i} \log \left[\frac{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}}{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})} + \sum\limits_{j \in \mathcal{D}, j \neq i} \frac{p(x_{2_j}|x_{1_i}, y_i)}{p(x_{2_j})}} \right] \right] \tag{25}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{p \sim \mathcal{P}_i} \log \left[1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} \sum_{j \in \mathcal{D}, j \neq i} \frac{p(x_{2_j} \mid x_{1_i}, y_i)}{p(x_{2_j})} \right] \right] \tag{26}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{p \sim \mathcal{P}_i} \log \left[1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} (N-1) \mathop{\mathbb{E}}_{j \in \mathcal{D}} \frac{p(x_{2_j} \mid x_{1_i}, y_i)}{p(x_{2_j})} \right] \right] \tag{27}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{p \sim \mathcal{P}_i} \log \left[1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} (N-1) \right] \right] \tag{28}$$

$$\geq \mathop{\mathbb{E}}_{i \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{p \sim \mathcal{P}_i} \log \left[\frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} N \right] \right] \tag{29}$$

$$= \log N - I(X_2; X_1, Y), \text{ using MI properties [38, Chapter 6.3.4]} \tag{30}$$

$$= \log N - I(X_2; Y | X_1) - I(X_2; X_1) \tag{31}$$

Therefore, by taking both sides of the contrastive loss to predict X_2 from X_1 and X_1 from X_2 we derive to $I(X_2; Y|X_1) + I(X_1; Y|X_2) + 2 \cdot I(X_2; X_1) \geq \log N - L_{\text{Con}}^{Opt}$. This trivially also holds for other ψ that obtain a worse(higher) L_{Con}. Simarly to InfoNCE, the bound becomes more accurate as N increases, while due to the the term $\frac{1}{|\mathcal{P}_i|}$ it is not affected by the number of positive pairs.