arXiv:2411.07335v2 [cs.LG] 7 Dec 2024

MULTIMODAL FUSION BALANCING THROUGH GAME-THEORETIC REGULARIZATION

Konstantinos Kontras
ESAT
KU Leuven

Thomas Strypsteen
ESAT
KU Leuven

Christos Chatzichristos
ESAT
KU Leuven

Paul Pu Liang
MIT Media Lab and MIT EECS

Matthew Blaschko
ESAT
KU Leuven

Maarten De Vos
ESAT
KU Leuven

ABSTRACT

Multimodal learning can complete the picture of information extraction by uncovering key dependencies between data sources. However, current systems fail to fully leverage multiple modalities for optimal performance. This has been attributed to modality competition, where modalities strive for training resources, leaving some underoptimized. We show that current balancing methods struggle to train multimodal models that surpass even simple baselines, such as ensembles. This raises the question: how can we ensure that all modalities in multimodal training are sufficiently trained, and that learning from new modalities consistently improves performance? This paper proposes the Multimodal Competition Regularizer (MCR), a new loss component inspired by mutual information (MI) decomposition designed to prevent the adverse effects of competition in multimodal training. Our key contributions are: 1) Introducing game-theoretic principles in multimodal learning, where each modality acts as a player competing to maximize its influence on the final outcome, enabling automatic balancing of the MI terms. 2) Refining lower and upper bounds for each MI term to enhance the extraction of task-relevant unique and shared information across modalities. 3) Suggesting latent space permutations for conditional MI estimation, significantly improving computational efficiency. MCR outperforms all previously suggested training strategies and is the first to consistently improve multimodal learning beyond the ensemble baseline, clearly demonstrating that combining modalities leads to significant performance gains on both synthetic and large real-world datasets.

1 INTRODUCTION

Exploiting multimodal data has made significant progress, with advances in generalizable representations and larger datasets enabling solutions to previously unattainable tasks [27; 29; 37; 43; 42; 44; 50; 53; 62]. However, studies indicate that multimodal data is often utilized suboptimally, underperforming compared to ensemble unimodal models or even the best single modality [55; 60]. The expectation that adding a modality should improve performance, assuming independent errors and above-chance predictive power [17], 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. Factors like noise levels, relationship complexity with the target, feature dimensionality, and data quality can cause one modality to fit faster than another. This imply that adding task-relevant information doesn't guarantee better performance, primarily due to complications during training. To address these issues, it's crucial to monitor each modality's contribution during training and apply corrective measures.

Several balancing strategies have been proposed to tackle this issue [5; 6; 9; 10; 21; 26; 28; 40; 41; 54; 55; 57; 60]. A central aspect of these methods is estimating each modality's contribution to the output. Most assume distributional independence between modalities on predicting the target, measuring contribution via unimodal performance [60; 26; 40; 6; 54]. Some methods bypass

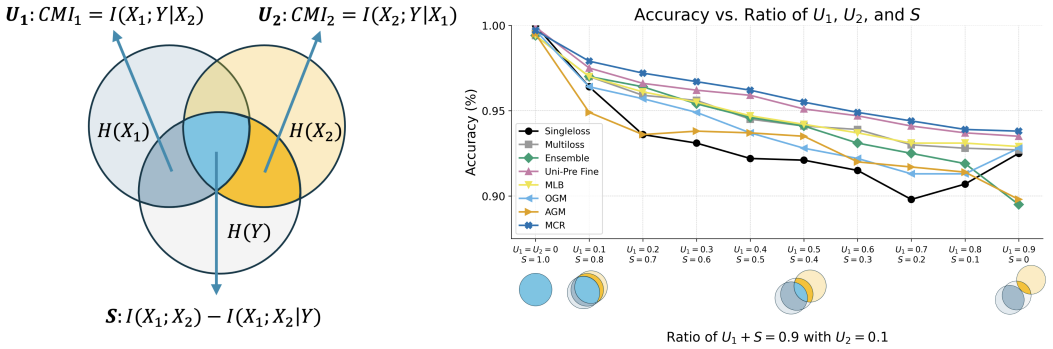

Figure 1: (Left) Illustration of the conditional mutual information (CMI) terms, $\text{CMI}_1 : I(X_1; Y \mid X_2)$ and $\text{CMI}_2 : I(X_2; Y \mid X_1)$, representing the unique contributions (U_1 and U_2) of each modality to the target. The shared task-relevant information (S) between the modalities is defined as $I(X_1; X_2) - I(X_1; X_2 \mid Y)$. (Right) Accuracy as a function of the ratio between the unique information (U_1) from modality X_1 and the shared information (S) between the modalities. Synthetic data are generated as $X_1 = N_1 + Y, X_2 = N_1 + Y$ where N_1, N_2 are independent noise for each modality. We consider S the percentage of the datapoints that both modalities have information about the label Y, U_1 and U_2 when only one has with the other modality equating to noise for those datapoints. In the experiment we keep U_2 constant while changing U_1 and S. As U_1 increases and S decreases, accuracy deteriorates, reflecting intensified multimodal competition. Among the various methods, including Singleloss, Multiloss, Ensemble, unimodally pretrained and finetuned encoders (Uni-Pre Fine), OGM [40], AGM [28], and MLB [26]our regularization method MCR demonstrates a slower decline in accuracy. For further detals prease refer to Section 4.1

this assumption by estimating influence based on prediction differences between original and perturbed inputs [28; 21; 10]. Perturbations can take various forms, such as zeroing values [28], adding Gaussian noise [10], or using task-specific augmentations [21; 31]. These methods aim to amplify a modality's influence by increasing the impact of perturbations on the output. However, this also makes the network more sensitive to these changes (e.g., noise), risks becoming overly reliant on the perturbations, and struggles to scale when multiple perturbations are required. Additionally, increasing the contribution of one modality can be achieved by overshadowing others, leading to an imbalance that undermines overall performance and makes the objective counterproductive.

Given these challenges, how can we design an efficient regularization method that addresses multimodal competition, ensuring balanced and effective learning across all modalities?

In this paper, we introduce the MULTIMODAL COMPETITION REGULARIZER (MCR), a loss function designed to promote the exploration of task-relevant information across all available modalities. By decomposing the joint mutual information (MI), we separately model shared and unique task-relevant information within the modalities. To efficiently capture the unique information from each modality, we employ a computationally inexpensive permutation-based approach. Our method maximizes the lower bounds of each MI term to encourage the network to learn both shared and unique information, while minimizing upper bounds on terms to suppress task-irrelevant information. We frame the problem in a game-theoretic setting, exploring strategies that involve both collaboration and competition among modalities to address the conflicting objectives that arise when increasing all modalities' contributions simultaneously. This approach allows their contributions to adapt dynamically, achieving balance during training.

We extensively evaluate MCR on synthetic datasets and several established real-world multimodal benchmarks, including action recognition on AVE [49] and UCF [45], emotion recognition on CREMA-D [4], human sentiment on CMU-MOSI [61], human emotions on CMU-MOSEI [63] and egocentric action recognition on Something-Something [15]. Our results demonstrate that MCR is the first balancing method to significantly improve supervised multimodal training over the ensemble baseline across a variety of datasets and models. Our key contributions are summarized as follows:

1. An analysis of multimodal competition, defining the error increase caused in multimodal training, while demonstrating in our results that most previous methods do no outperform simple baselines, such as unimodal ensembles.
2. A novel multimodal training strategy, MCR, designed to regularize multimodal competition, which includes:
   - Defining lower and upper bounds of the MI terms, encouraging the exploration of information across all modalities.
   - Introducing a game-theoretic perspective where modalities form the players that compete for training resources, assisting the regularization through balancing the corresponding MI terms.
   - Suggesting latent-space perturbations as an efficient way to estimate the lower bound of the CMI reducing the computational cost of multiple forward passes.

2 Analysis of Multimodal Competition

Consider a dataset of N i.i.d. datapoints sampled from a distribution $\mathcal{D}$, where each datapoint consists of M modalities $X = (X_1, .., X_M)$ and a target Y. Our objective is to learn a parameterized function $f : X; \theta \rightarrow Y$. We define the unimodal encoder for each modality as $f_m : X_m; \theta_m \rightarrow Z_m$, which encodes input X_m into a latent representation Z_m. The fusion network $f_c : [Z_1, \ldots, Z_M]; \theta_c \rightarrow Y$ takes these latent representations and predicts the target Y, as does independently on each modality the unimodal task head $f_{c_m} : Z_m; \theta_{c_m} \rightarrow Y_m$. The families of unimodal and multimodal models are defined as follows:

$$\text{Unimodal models } \mathcal{F}_{u_m} : f_{u_m}\left(X_m; \theta_{u_m}\right) = f_{c_m}\left(f_m\left(X_m; \theta_m\right); \theta_{c_m}\right), \text{ for } m = [1, .., M]. \quad (1)$$

$$\text{Multimodal models } \mathcal{F} : \quad f(X; \theta) = f_c\left(f_1\left(X_1; \theta_1\right), \ldots, f_M\left(X_M; \theta_M\right); \theta_c\right). \quad (2)$$

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

The limitation of supervised multimodal training

In supervised learning with multimodal inputs X_1 and X_2, the objective is to learn representations $Z_1 = f_1(X_1; \theta_1)$ and $Z_2 = f_2(X_2; \theta_2)$ that, when combined through a fusion model $f_c(Z_1, Z_2)$, accurately predict a target Y. This can be expressed as minimizing the expected task loss or by maximizing the mutual information between the fused representation and the target:

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

During training, the model often undertrains the weaker modality, prioritizing the more accessible or higher-quality one. Such over-reliance on one modality limits the model's ability to effectively utilize all available information. Consequently, the mutual information becomes dominated by the stronger modality, resulting in $I(f_c(Z_1, Z_2); Y) \approx I(Z_1; Y)$. In this scenario, the conditional mutual information (CMI) $I(Z_2; Y \mid Z_1) \approx 0$, indicating that once the model learns from Z_1, the information from Z_2 contributes little to predicting Y. Appendix A.1 presents a brief experiment demonstrating this phenomenon.

This phenomenon is not exclusive to multimodal learning. In single-modality feature learning, models often fit faster to dominant features, neglecting others that could enhance generalization. Regularization techniques like L1/L2 penalties and dropout [38; 46] were introduced to encourage balanced feature use and promote the learning of diverse patterns. While these techniques have become the standard in training models, their adaptation to multimodal learning has proven more challenging. For instance, Xiao et al. [58] attempted to apply dropout individually to each modality, but subsequent research [40] demonstrated that this approach is still limited in addressing the core issue. The central problem lies in effectively regulating how modalities interact and compete, which remains an open question in the field.

Multimodal competition

Multimodal competition occurs when the network primarily optimizes for one modality, leading to a decline in generalization. That modality reduces training error by overfitting to the data, limiting

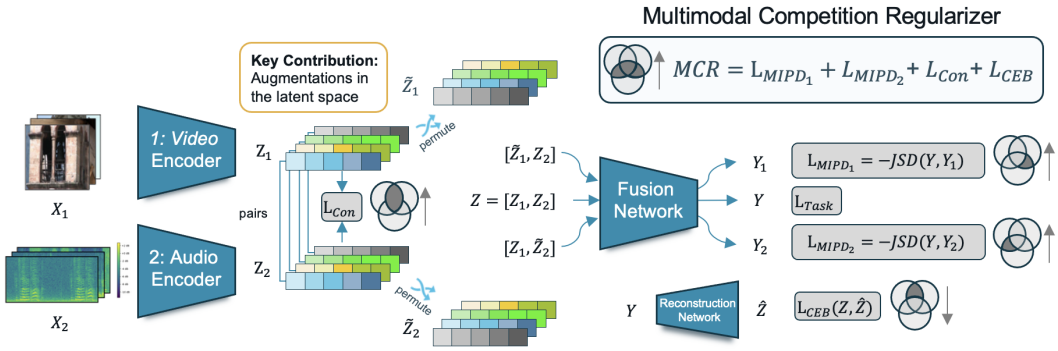

Figure 2: **Multimodal Competition Regularizer (MCR):** The diagram illustrates the MCR framework used to mitigate competition between modalities in multimodal learning. Raw data (X_1 and X_2) are fed into the respective encoders, generating latent representations (Z_1 and Z_2). Each unimodal representation is permuted to create $\tilde{Z}_1$ and $\tilde{Z}_2$, and the three combinations are fed into the Fusion Network, producing predicted outputs (Y, Y_1, Y_2). The comparison between the predictions allows for an assessment of unimodal contribution. For instance, if $Y \approx Y_1$, then the disturbance of Z_1 had no impact on the result, indicating that the model relies solely on modality X_2, making the contribution of X_1 negligible. The MCR loss consists of three components: $\mathcal{L}_{MIPD1}$ and $\mathcal{L}MIPD_2$ maximize the Jensen-Shannon divergence (JSD) between the task output and predictions with permuted modalities, ensuring that each modality has the maximum information contribution. $\mathcal{L}_{\text{Con}}$ encourages alignment between the representations of the modalities, promoting to capture the shared information. $\mathcal{L}_{\text{CEB}}$ applies the conditional entropy bottleneck to regulate the mutual information between the combined representation Z and the reconstructed output $\hat{Z}$. This framework aims to balance the contributions of each modality while minimizing competition, allowing for more robust multimodal representations.

gradient feedback to the other modalities. Huang et al. [20] demonstrate 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 $\Gamma = \sup\langle X, W\rangle$ of each input modalities and the weights of the output layer. If a modality dominate the competition it causes the other modality to maintain its initial (before training) correlation levels. We proceed this analysis by introducing the generalization error ϵ resulting from the effects of multimodal competition.

Definition of Multimodal Competition Error (MCE): Let Γ_1 and Γ_2 represent the correlations of modalities X_1 and X_2 with the output. If, at the end of 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 ($\mathbb{E}[e_1 e_2] = 0$, 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:

$$\epsilon \geq R(f^*) - R_{\text{emp}}(f^*) \text{ and } \epsilon \leq R(f) - R_{\text{emp}}(f) \tag{4}$$

where $R(f) = \mathbb{E}[L(y, f(x))]$ and $R_{\text{emp}}(f)$ denote the generalization and empirical model risks.

A higher ϵ indicates a stronger effect of multimodal competition, implying that the dominance of one modality significantly impacts the model's generalization. In the following section, we introduce MCR as method to minimize ϵ by maximizing each modality's correlation with the output.

3 Multimodal Competition Regularizer

In this section, we introduce $\mathcal{L}_{\text{MCR}}$, a set of targeted loss components designed to directly address and mitigate the challenges of multimodal competition. These losses are derived by framing the multimodal learning task as optimizing the mutual information between the target Y and the joint distribution of unimodal input features $X = (X_1, X_2)$. The method is focusing on two input modalities, though an extension to M modalities is presented in Appendix A.5. The motivation behind these loss components stems from this formulation. To clearly define each loss, we first decompose the mutual information into distinct terms:

$$I(X_1;X_2;Y) = \underbrace{I(X_1;Y \mid X_2) + I(X_2;Y \mid X_1)}_{\substack{\text{Task-Relevant Unique Information} \\ \text{of each modality} \sim \mathcal{L}_{\text{MIPD}}}} + \underbrace{I(X_1;X_2)}_{\substack{\text{Shared Information} \\ \mathcal{L}_{\text{Con}}}} - \underbrace{I(X_1;X_2 \mid Y)}_{\substack{\text{Task-Irrelevant} \\ \text{Shared Information} \sim \mathcal{L}_{\text{CEB}}}} . \tag{5}$$

This decomposition is illustrated by the Venn diagram in Figure 1. The first two terms, the CMIs $I(X_1;Y \mid X_2)$ and $I(X_2;Y \mid X_1)$, represent the unique information each modality holds about the target that is not shared with the other modality. Maximizing these terms encourages the model to extract modality-specific task-relevant features. The third term, $I(X_1;X_2)$, quantifies shared information between the modalities. Maximizing it encourages the model to align the representations and effectively utilize the common information between [22; 39; 44].

The third term, $I(X_1;X_2)$, quantifies how much information the two modalities share. Maximizing this term encourages the model to align the representations and effectively utilize the common information between them [22; 39; 44]. The third term, $I(X_1;X_2)$, quantifies the shared information between the modalities. Maximizing this term encourages to explore common information and aligns the representations, similar to how contrastive learning exploits multi-view redundancy to enhance representation quality [22; 39; 44]. The final term, $I(X_1;X_2 \mid Y)$, represents task-irrelevant shared information. It penalizes the shared information between modalities but not useful for predicting the target. For each of these terms, we introduce a corresponding loss component. The total regularization loss is defined as follows:

$$\text{Regularization Loss:} \quad \mathcal{L}_{\text{MCR}} = \mathcal{L}_{\text{Con}} + \mathcal{L}_{\text{CEB}} + \mathcal{L}_{\text{MIPD}}, \text{with} \tag{6}$$

$$\text{Competition Terms:} \quad \mathcal{L}_{\text{MIPD}} = (\underset{\theta_1}{\mathcal{L}_{\text{MIPD}_1}} - \underset{\theta_1}{\mathcal{L}_{\text{MIPD}_2}}) + (\underset{\theta_2}{\mathcal{L}_{\text{MIPD}_2}} - \underset{\theta_2}{\mathcal{L}_{\text{MIPD}_1}}). \tag{7}$$

The proposed regularizer consists of three key losses: $\mathcal{L}_{\text{Con}}$, a contrastive loss that captures shared information between modalities, $\mathcal{L}_{\text{CEB}}$, which filters out task-irrelevant information, focusing on relevant features for the downstream task, and $\mathcal{L}_{\text{MIPD}}$, which measures each modality's affection to the multimodal prediction by assessing output variations under perturbations. The latter has contradicting terms since increasing the importance of one modality decreases the importance of the other, and hence we model it as a game where each modality encoder (with parameters θ_1, θ_2) competes to increase the importance of its own modality while decreasing the other. In the remainder of this section, we will examine each component individually and further elaborate on framing multimodal competition using game theory.

3.1 Approximating MI Terms

To approximate each CMI and capture the unique contribution of each modality, the Mutual Information Perturbed Difference (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.

The perturbed version of modality X_1 is denoted as $\tilde{X}1$. We use permutations as our perturbation method, which will be detailed further in Section 3.3.

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

$$\text{MIPD}(X_1;Y \mid X_2) = I(X_1;Y \mid X_2) - I(\tilde{X}_1;Y \mid X_2) \leq I(X_1;Y \mid X_2). \tag{8}$$

Using the entropy interpretation of the mutual information terms, each CMI can be expressed as the difference of the log probabilities with and without the perturbations:

$$\begin{aligned} \text{MIPD}(X_1;Y \mid X_2) &= H(Y \mid X_2, \tilde{X}_1) - H(Y \mid X_2, X_1) \\ &= \mathbb{E}_{\substack{y \sim p(y) \\ x_1, x_2 \sim p(x_1, x_2, y)}} \left[-\mathbb{E}_{\tilde{x}_1 \sim p(x_1)} \left[\log p(y \mid x_2, \tilde{x}_1) \right] + \log p(y \mid x_2, x_1) \right]. \end{aligned} \tag{9}$$

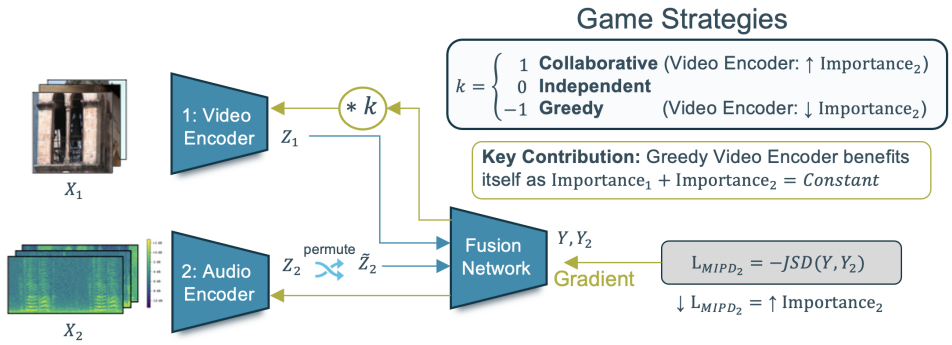

Figure 3: In this figure, we highlight a specific part of our training process to demonstrate how different competition strategies between modalities are applied. The gradient multiplier adjusts the video encoder's behavior toward Audio Importance (Importance_2). When $k = 1$, the video encoder collaborates to boost Importance_2. At $k = 0$, it remains neutral. At $k = -1$, it competes by reducing Importance_2 to increase its own (Importance_1). This is based on the principle that Total Importance ($\text{Importance}_2 + \text{Importance}_1$) remains constant.

Moreover, we employ the Jensen-Shannon divergence (JSD) [32], which is symmetrically bounded, to avoid training instabilities leading to the following objective:

$$\mathcal{L}_{\text{MIPD}_{\text{X}_1}} = -\,\text{MIPD}(X_1; Y \mid X_2) = -\mathbb{E}_{\substack{y \sim p(y) \\ x_1, x_2 \sim p(x_1, x_2, y) \\ \tilde{x}_1 \sim p(x_1)}} \left[\text{JSD}(P(y \mid x_2, x_1), P(y \mid x_2, \tilde{x}_1))\right]. \tag{10}$$

Next, the $I(X_1; X_2)$ MI term measures how much information the two modalities share with each other. It the common patterns between the two, ensuring that the model recognizes and aligns these shared aspects. We exploit the available label information employing the supervised contrastive loss [24] $\mathcal{L}_{\text{Con}}$. The term is defined as follows:

$$\mathcal{L}_{\text{Con}}(X_1, X_2) = \mathbb{E}_{\substack{x_1, y \sim p(x_1, y) \\ x_2^+ \sim p(x_2 \mid y) \\ x_2^- \sim p(x_2 \mid \neg y)}} \left[\log \frac{\psi(x_1, x_2^+)}{\sum_k \psi(x_1, x_2^-)}\right], \tag{11}$$

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

$$I(X_2; Y|X_1) + I(X_1; Y|X_2) + 2 \cdot I(X_2; X_1) \geq \log N - \mathcal{L}_{\text{Con}}^{Opt} \tag{12}$$

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

Finally, $I(X_1; X_2 \mid Y)$ 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 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 into the latent space, effectively filtering out irrelevant content:

$$\mathcal{L}_{\text{CEB}}(X_1; X_2) = \mathbb{E}_{x_1, x_2, y \sim p(x_1, x_2, y)} \left\|[f_1(x_1), f_2(x_2)] - h(p(y \mid x_1, x_2))\right\|^2 \tag{13}$$

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.

3.2 The Game of Multimodal Fusion

In this section, we present the game-theoretic approach used to balance the terms of the proposed multimodal training regularizer. The key insight is that the total contribution from all modalities, expressed through modality importance, remains constant. This implies that when one modality's

contribution increases significantly, it may diminish the contributions of other modalities. This allows to frame multimodal competition as a constant-sum game where each modality acts as a player. The $\mathcal{L}_{\text{MIPD}}$ corresponding to each modality represents the payoff function for that modality, where each modality aims to maximize its own contribution to the overall output. The players take concurrent actions, as each modality is updated during every optimization step. We constrain the players' actions by enforcing a fixed strategy throughout the entire training process. We explore three strategic actions for the modalities:

- **Collaborative Actions:** All modalities work together to increase each other's contributions. The $\mathcal{L}_{\text{MIPD}}$ terms are applied across all parameters, resulting in a joint minimization, $\min_{\theta} \mathcal{L}_{\text{MIPD}}$
- **Independent Actions:** Each modality focuses on maximizing its own contribution by optimizing its respective $\mathcal{L}_{\text{MIPD}}$ term, without regard to the contributions of other modalities, leading to $\min_{\theta_i} \mathcal{L}_{\text{MIPD}_{X_i}}$.
- **Greedy Actions:** 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. In the constant-sum game, reducing the influence of other modalities enhances its own contribution, resulting in a min-max strategy, $\min_{\theta_i} \max_{\theta_{\neg i}} \mathcal{L}_{\text{MIPD}_{X_i}}$ [1].

In Figure 3, we provide an example of the choices the video modality can make, illustrating how it can either assist, ignore, or diminish the importance of the audio modality. By default, we adopt the greedy actions strategy, unless otherwise specified.

3.3 Perturbations

To measure the importance of modality X_1, we observe the difference in its predictive output when presented with $\{\tilde{X}_1, X_2\}$ compared to the normal input $\{X_1, X_2\}$.

Previous works have explored various perturbation techniques. Gat et al. [10] used additive noise to estimate functional entropy by maximizing output variance. Maximizing this variance can increase sensitivity to noise, undermining the goals of Lipschitz smoothness [1] and adversarial robustness [47]. Ji et al. [21] and Liang et al. [31] proposed task-specific augmentations, that may not be available in most tasks and modalities, while their method relies heavily on the success of these augmentations. Li et al. [28] employed zero-masking to estimate Shapley values and trained directly to maximize them. However, Shapley values can produce arbitrary or unreliable estimates, particularly in high-dimensional settings [35]. Additionally, each of these methods perturbs the input, requiring extra forward passes and proportionally increasing computational and memory demands.

To address the limitations of the suggested perturbations and face their computational need, we propose as perturbation to use within-batch permutations $\sigma_e \sim \text{Uniform}(\mathcal{P})$, where $\mathcal{P}$ represents the set of all possible permutations [11]. Permutations differ from previous methods by applying the transformation directly to the latent space, eliminating the need for multiple forward passes through the unimodal encoders, thereby reducing both computational and memory overhead. To analyze the effects of permutations, we examine two cases separately: when the label of the original data matches the label of the permuted pair, and when it does not:

1. **Matching Labels:** The normal and perturbed latent vectors represent the same class, allowing us to test if changes in task-irrelevant information impact the output, resembling the optimal augmentation described in Liang et al. [31]. Maximizing their output difference can help the model spread the datapoint representations within the class boundaries.
2. **Non-Matching Labels:** In this case, the normal and perturbed latent vectors belong to different classes. Maximizing their output difference can promote better class discrimination, refining the decision boundaries and reinforcing model's robustness and calibration on out-of-distribution cases.

A detailed analysis of the computational benefits of applying perturbations to the latent space is provided in Appendix A.6, along with a comparison of perturbations in Appendix C.1.

[1] 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 $X_1, X_2, \ldots, X_n$, output 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 epoch **do**
  for each batch $(X_1, .., X_n, Y_t)$ sampled from D **do**
    Estimate $Y, Z_1, .., Z_n = f(X_1, .., X_n)$
    Compute $L_{task}(f(X_1, .., X_n), Y_t)$ and $L_{task}^{uni} = \lambda_{\text{uni}} \sum_{i=1}^{M} L_{task}(f_i^u(X_i), Y_t)$
    Compute the L_{Con} with Eq. 11 and L_{CEB} with Eq. 13
    Sample P_e permutations and compute permuted pairs on the latent space Z
    Pass each pair through the fusion model f_c to get predictions $\tilde{Y}_i$ with modality i permuted
    Compute $L_{\text{MIPD}} = \lambda_{\text{M}} \sum_{i=1}^{M} L_{\text{MIPD}_{x_i}}$ using Eq. 10 and the corresponding strategy
    Compute L_{MCR} from Eq. 6
    Update with $L_{total} = L_{task} + L_{uni} + L_{\text{MCR}}$
  end for
end for

4 Experiments

We perform extensive experiments on synthetic and real-world multimodal datasets to evaluate the effectiveness of our proposed method, MCR, against various baselines and previous methods:

Unimodal Training: Separate training for each modality.

Ensemble: Combines unimodal predictions without any further training.

Joint Training: Trains all modalities under a single-loss objective, without any balancing.

Multi-Loss [52]: Incorporates additional unimodal task losses alongside the joint objective.

MMCosine [59]: Equalizes modality contributions by standardizing features and weights.

MSLR [60]: Adapts unimodal encoders' learning rates based on validation performance.

PMR [6]: Similar to Multi-Loss, uses prototype classifiers and adjusts unimodal loss weights.

OGM [40]: Extracts unimodal performance from the last layer, compares improvements, and modulates gradients by adjusting each encoder's learning rate.

AGM [28]: Uses zero-masking Shapley values as predictor for the unimodal task losses and modulate them by adjusting their coefficients.

MLB [26]: Combines Multi-Loss for enhanced unimodal performance assessment and adjusts each encoder's learning rate.

Uni-Pre Frozen: Uses pre-trained unimodal encoders with frozen weights during joint training.

Uni-Pre Finetuned: Employs pre-trained unimodal encoders, fine-tuned during joint training.

4.1 Synthetic Dataset

We generate an artificial scenario where the mutual information between the label and the modalities varies, demonstrating the phenomenon of modality competition. Although many factors can contribute to this effect, a comprehensive exploration is beyond the scope of this work. Instead, we focus on the imbalance in modality informativeness to highlight the motivation behind our approach.

Data: We generate task-irrelevant information for each modality by sampling $N_1, N_2 \sim \mathcal{N}(0, I)$ and the 5-class label Y from a uniform distribution $Y \sim \text{Uniform}(5)$. Using fixed transformations, similar to Liang et al.[31], each modality is converted into a high-dimensional vector. We relate both modalities to the label through a linear relationship: $X_1 = N_1 + Y$ and $X_2 = N_2 + Y$. Data points are distributed such that either both modalities contain label information (Shared Information) or only one of the modalities (Unique Information). We vary the percentage of data points with shared and unique information to analyze how models perform 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 for all methods a performance drop. 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 have either been identified as exhibiting significant imbalance among modalities or serve as standard multimodal benchmarks:

CREMA-D [4]: An emotion recognition dataset featuring 91 actors expressing 6 distinct emotions.
AVE [49]: A collection of videos with temporally aligned audio-visual events across 28 categories.
UCF [45]: An action recognition dataset consisting of real-life YouTube videos.
CMU-MOSEI [63]: Large-scale multimodal sentiment analysis dataset with 23k monologue clips.
CMU-MOSI [61]: Multimodal sentiment analysis dataset with over 2k YouTube video clips.
Something-Something (V2) [15]: Contains 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 allows us to demonstrate 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 as alongside with state-of-the-art models like Conformer [14] Swin-TF [33]. Detailed model configurations for each dataset are provided in Appendix A.2. These choices aim to bridge the gap between theoretical work and practical application.

Table 1: This table compares the proposed MCR method with several baselines across the additional datasets: CMU-MOSEI, CMU-MOSI, and Something-Something. Results are shown for different modality combinations (e.g., V-A-T: Video-Audio-Text), using Transformer and Swin-TF backbones. MCR consistently outperforms other methods across all configurations.

	Transformer				Swin-TF
	MOSEI		MOSI		Something-Something
Method	V-T	V-A-T	V-T	V-A-T	V-OF
Unimodals	V: 63.6 A: 64.3 T: 79.9		V: 53.5 A: 54.9 T: 70.7		V:61.6 OF:50.6
Ensemble	77.4	76.6	69.8	43.3	64.8
Joint Training	80.5	81.2	74.7	73.9	56.3
Multi-Loss	80.8	79.9	75.3	72.0	59.6
OGM [40]	80.6	-	75.2	-	58.3
AGM [28]	79.5	80.6	75.3	75.2	56.9
MLB [26]	80.8	81.1	75.0	74.5	61.3
Uni-Pre Frozen	80.5	79.8	70.6	70.6	64.2
Uni-Pre Finetuned	80.7	80.1	73.8	71.2	61.9
MCR	**81.2**	**81.6**	**76.1**	**77.9**	**65.0**

Results: In Figure 4 and in Table 1 we present the accuracy comparison among the different baseline methods on our datasets. In the figure we provide accuracy as a comparison with that of Ensemble to highlight the difference between multimodal and unimodal training capabilities. We use both backbone encoders where we identified good SSL pretrained fit models: Conformer [14](in orange) and ResNet (in blue). Positive values indicate an improvement over the ensemble baseline, while negative values signify a decline in performance. Our main observations are:

a) Most previous methods and baselines fail to outperform the Ensemble, often showing negative accuracy differences, particularly on the CREMA-D and AVE datasets. These methods do not adequately address the modality competition problem, leading to a worse generalization error struggling to leverage the given multimodal data, regardless of the chosen backbone model.

b) Our method, MCR, stands out as the only approach that consistently achieves better across all datasets and backbone models. Notably, MCR outperforms consistently the Ensemble baseline across all models and datasets, demonstrating its effectiveness in balancing modality contributions under several different configurations.

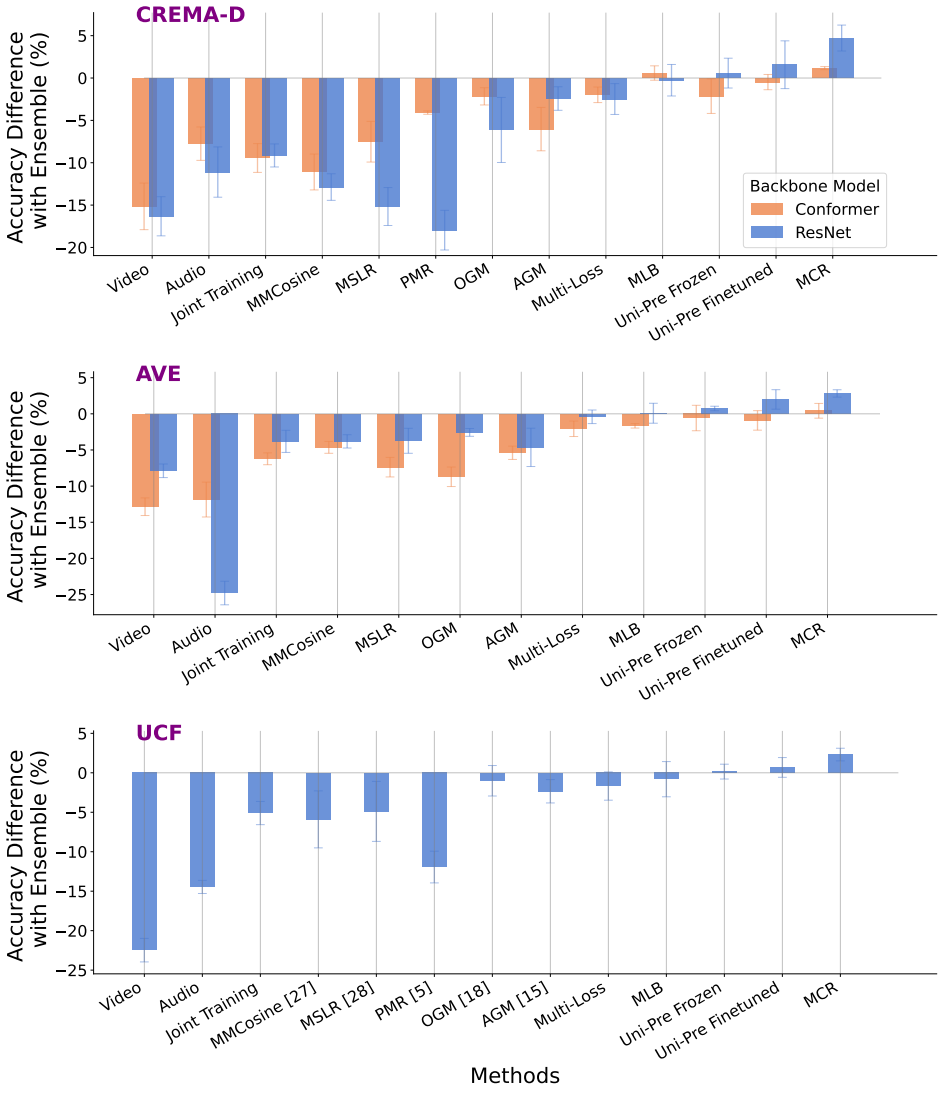

Figure 4: Accuracy(%) difference compared to the Ensemble baseline for various methods across three datasets: CREMA-D, AVE, and UCF101. Results are shown for two backbone models: Conformer (orange) and ResNet (blue). Positive values indicate an improvement over the Ensemble, while negative values represent a performance drop. Most existing methods fail to outperform the Ensemble, which does not utilize any multimodal learning. In contrast, our proposed method, MCR, consistently achieves positive accuracy differences, surpassing both the previous suggested methods and our own defined baselines.

4.3 Ablation Study

Strategy: 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 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 for different game strategies: Collaborative, Independent, and Greedy, using both ResNet and Conformer backbones.

Model Setting	Collaborative $\min_{\theta} \mathcal{L}_{\text{MIPD}}$	Independent $\min_{\theta_i} \mathcal{L}_{\text{MIPD}_{\text{X}_\text{i}}}$	Greedy $\min_{\theta_i} \max_{\theta_{\neg i}} \mathcal{L}_{\text{MIPD}_{\text{X}_\text{i}}}$
ResNet + MCR	73.4$\pm$3.0	76.0$\pm$2.0	**76.2$\pm$1.7**
Conformer + MCR	82.9$\pm$0.7	82.9$\pm$1.4	**85.7$\pm$0.2**

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 L_{MIPD} and L_{Con} is yielding consistent improvement. Exploiting them separately leads to smaller improvement for L_{Con} and even to a decline for L_{MIPD}. suggests that alignment in the latent space between the modalities is necessary for the permutations to be effective.
2. The L_{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 L_{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: L_{MIPD}, L_{Con}, and L_{CEB}. Results indicate that combining L_{MIPD} and L_{Con} is crucial for improvement, while L_{CEB} does not benefit all models.

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

4.4 Analysis of Multimodal Error

In this section, we investigate the specific data points that MCR prioritizes during training. The framework is designed to balance and enhance the independent contributions of each modality through the CMI terms, while strengthening their shared contributions via the contrastive loss. However, no specific term explicitly addresses the synergetic, emergent information that arises from the interplay between modalities. The underlying expectation is that encouraging modalities to improve concurrently would naturally lead to such synergetic behavior. Here, we conduct a post-hoc error analysis to examine the results and better understand the benefits that MCR provides.

We analyze a matrix comparing the unimodal wrong and correct predictions with those from multimodal models trained using various methods, including Ensemble, Joint Training, AGM, MLB, and MCR. As shown in Figure 5, a consistent pattern emerges across different dataset/model pairs. MCR outperforms previous methods in cases where at least one modality independently makes the correct prediction, with the ResNet-based models even surpassing the Ensemble in this category. However, we observe that while MLB and AGM struggle to retain unimodal information effectively, they consistently perform better at discovering new emergent information. MCR improves upon this

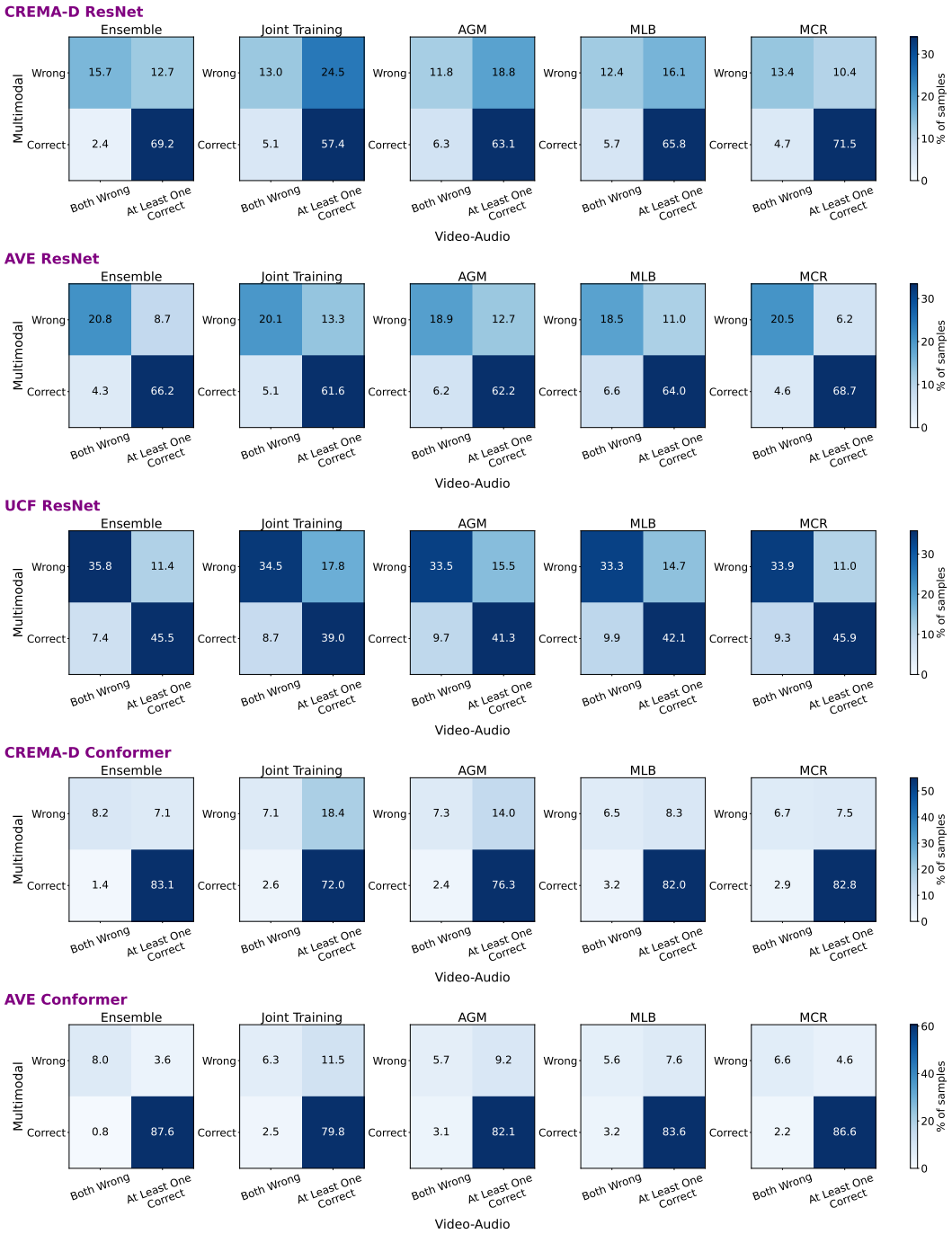

Figure 5: Error comparison matrices across pairs of datasets and models, comparing unimodal predictions with multimodal models trained using various methods, including Ensemble, Joint Training, AGM, MLB, and MCR. Each column of the confusion matrix represents cases where both unimodal predictions are 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, with ResNet-based models surpassing even the Ensemble in this category. Additionally, MLB demonstrates stronger performance in discovering emergent information, highlighting a current limitation of MCR in this aspect.

in certain instances by approaching the best-performing baselines but it generally shows a decline in that category. This analysis is encouraging for two reasons: it highlights a significant percentage of unimodal information that is not adequately leveraged by the multimodal model, and it indicates potential for further improving the discovery of emergent information, both of which are crucial for achieving performance gains that justify the effort required to collect, process, and compute the additional multimodal data.

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.

Looking ahead, further advancements could involve incorporating additional terms to explicitly encourage synergetic interactions between modalities, alongside an exploration of game-theoretic modeling to enable more flexible and adaptive strategies. Future work may focus on refining these strategies, enabling individualized and adaptive decisions for each modality, and tightening mutual information bounds to unlock even greater performance improvements.

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] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[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] 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.

[24] 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.

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

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

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

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

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. *arXiv preprint arXiv:2402.16318*, 2024.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

A Appendix / supplemental 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 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 6, 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 the any information available in the video modality and performs significantly worse than an ensemble of the unimodally trained models.

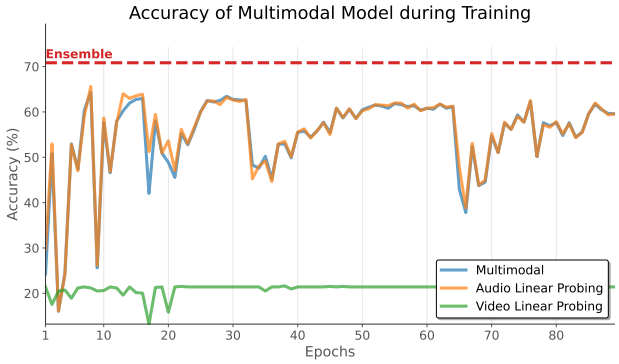

Figure 6: Accuracy of the multimodal model on the CREMA-D dataset across training epochs, showing the performance of the full multimodal model (blue) and individual modality linear probing for audio (orange) and video (green). The dashed red line represents the accuracy of a unimodal ensemble model, highlighting how the model's over-reliance on the audio modality negatively impacts the utilization of the video modality

A.2 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 [40; 28; 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. [13], incorporating audio signals sampled at 16 kHz and utilizing pre-calculated facial features. Unlike previous approaches [40; 28; 6], our dataset division follows Goncalves et al. [13], excluding actor overlap between training, validation, and test sets. We report standard deviation (std) across folds for consistency.

AVE [49]: 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 [45]: 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 [45], reporting the std across these folds.

Something-Something (V2)[15]: 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. [43], 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. [43]. Results are reported based on a single run.

CMU-MOSI[61] and -MOSEI[63]: 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 [50] 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 [30].

A.3 Models: Backbone Unimodal Encoders

In line with previous research [28; 40; 6; 59], our initial experiments adopt ResNet-18 [18] 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 [51] similar to previous works [31; 30].

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 [14] on CREMA-D, we deploy the first 12 layers of the Wav2Vec2 [3; 56] 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 [64] and afterwards the facial features of every available frame exploiting EfficientNet-B2 [48] as a frozen feature descriptor. The extracted audio features and pre-calculated facial features are further refined using a 5-layer Conformer [16], initialized from scratch. 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 [23], and for the audio branch, the HuBERT [19] model pretrained on Audioset [12]. Ensure 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. [43], which highlight the importance of modality-specific processing in multimodal tasks. For both video and optical flow data, we adopt the Swin Transformer [33] as backbone to each modality. This state-of-the-art architecture excels in capturing hierarchical and spatiotemporal features through its shifted window attention mechanism.

A.4 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 by one magnitude lower.

All models are optimized using Adam [25] with a cosine learning rate scheduler and a steady warm-up phase, except for the Something-Something dataset, where we use Adaw [34]. 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. 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 consistently $lr = 1\text{e}{-4}$ 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, 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.

A.5 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:

$$I(X_1, .., X_M; Y) = \sum_{\mathcal{S} \subseteq \{X_1, .., X_M\}, \mathcal{S} \neq \emptyset} I(\mathcal{S}; Y \mid \{X_1, \ldots, X_M\} \setminus \mathcal{S}) + I(X_1, .., X_M) - I(X_1, .., X_M \mid Y), \quad (14)$$

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.6 COMPUTATIONAL SPEED AND MEMORY ANALYSIS

The computational load imposed by any sample that requires additional pass is 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 overall computational complexity is $\mathcal{O}(M * N * (cost_{enc} + cost_c))$, where M is the times we draw noisy samples and N is the batch size. If we now use permutation samples which can be directly drawn from the latent space, the additional computational complexity is reduced to $\mathcal{O}(N * cost_{enc} + M * N * cost_c))$. In most state-of-the-art models each modality encoders 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:

$$\mathcal{L}_{\text{Con}}(X_1, X_2) = \sum_{i \in \mathbb{I}} \frac{-1}{|\mathcal{P}_i|} \sum_{k \in \mathcal{P}_i} \left[\log \frac{\psi(x_{1_i}, x_{2_k})}{\sum_{j \in \mathbb{I}} \psi(x_{1_i}, x_{2_j})} \right] \tag{15}$$

where $\mathcal{P}_i = \{p \in \mathbb{I} \mid y_p = y_i\}$. In supervised contrastive learning, the presence of multiple positive samples turns this into a multi-label problem, unlike traditional NCE methods [39], 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 $p(x2_p \mid x_{1_i}, y_i)$ while sampling the remaining points from the proposal distribution $p(x_{2_l})$. This approach mirrors the technique used in InfoNCE [39] and leads to the following derivation:

$$p(d = p|X_2, x_{1_i}, y_i) = \frac{p(x_{2_p} \mid x_{1_i}, y_i) \prod_{l \in \mathcal{I}, l \neq p} p(x_{2_l})}{\sum_{j \in \mathcal{I}} p(x_{2_j} \mid x_{1_i}, y_i) \prod_{l \in \mathcal{I}, l \neq j} p(x_{2_l})} \tag{16}$$

$$= \frac{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}}{\sum_{j \in \mathcal{I}} \frac{p(x_{2_j}|x_{1_i}, y_i)}{p(x_{2_j})}} \tag{17}$$

The optimal value for the critic function ψ in Equation 17 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:

$$L_{\text{Con}}^{Opt} = -\mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[\mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[\frac{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}}{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})} + \sum_{j \in \mathcal{I}, j \neq i} \frac{p(x_{2_j}|x_{1_i}, y_i)}{p(x_{2_j})}} \right] \right] \tag{18}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[\mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} \sum_{j \in \mathcal{I}, j \neq i} \frac{p(x_{2_j} \mid x_{1_i}, y_i)}{p(x_{2_j})} \right] \right] \tag{19}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[\mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} (N-1) \mathop{\mathbb{E}}_{j \in \mathcal{I}} \frac{p(x_{2_j} \mid x_{1_i}, y_i)}{p(x_{2_j})} \right] \right] \tag{20}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[\mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} (N-1) \right] \right] \tag{21}$$

$$\geq \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[\mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[\frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} N \right] \right] \tag{22}$$

$$= \log N - I(X_2; X_1, Y), \text{ using MI properties [36, Chapter 6.3.4]} \tag{23}$$

$$= \log N - I(X_2; Y|X_1) - I(X_2; X_1) \tag{24}$$

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.

C ADDITIONAL ABLATION STUDIES

C.1 COMPARING PERTURBATION METHODS

In this work, we explored 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.

Table 4 presents 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. The methods include:

1. Noise in the Input Space: Adding noise directly to the input features of each modality, simulating realistic data corruption. For its implementation we follow Gat et al. [10].
2. Shapley Input-Space Perturbations: Following the approach of Li et al. [28], Shapley zero-induced values are used to determine the importance of input modalities.
3. Noise in the Latent Space: Applying noise to the latent representations, encouraging robustness at the feature extraction level.
4. Within-Batch Permutations in the Latent Space: Permuting data points within the batch to disrupt alignment.

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.6, 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.

Table 4: Ablation study comparing different perturbation methods for MCR across multiple datasets. The table shows the performance of MCR when combined with various perturbation techniques, including noise in the input and latent space, Shapley values in the input space, and within-batch permutations.

	Datasets				
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$0.8	80.9	76.1
MCR with Shapley Input-Space	73.6$\pm$1.5	72.6$\pm$0.9	**55.5$\pm$0.6**	81.2	**76.2**
MCR with Noise Latent-Space	73.6$\pm$1.1	72.6$\pm$0.4	54.5$\pm$0.7	**81.3**	75.5
MCR with Permutations	**76.1$\pm$1.1**	**73.3$\pm$0.5**	55.2$\pm$1.8	81.2	**76.2**