Learning Invariant Modality Representation for Robust Multimodal Learning from a Causal Inference Perspective

Sijie Mai* **Shiqin Han**
School of Computer Science, South China Normal University
{sijiemai,20222121019}@m.scnu.edu.cn

Abstract

Multimodal affective computing aims to predict humans' sentiment, emotion, intention, and opinion using language, acoustic, and visual modalities. However, current models often learn spurious correlations that harm generalization under distribution shifts or noisy modalities. To address this, we propose a causal modality-invariant representation (CmIR) learning framework for robust multimodal learning. At its core, we introduce a theoretically grounded disentanglement method that separates each modality into 'causal invariant representation' and 'environment-specific spurious representation' from a causal inference perspective. CmIR ensures that the learned invariant representations retain stable predictive relationships with labels across different environments while preserving sufficient information from the raw inputs via invariance constraint, mutual information constraint, and reconstruction constraint. Experiments across multiple multimodal benchmarks demonstrate that CmIR achieves state-of-the-art performance. CmIR particularly excels on out-of-distribution data and noisy data, confirming its robustness and generalizability.

1 Introduction

Multimodal affective computing (MAC) aims to integrate information from language, acoustic, and visual modalities to predict high-level semantics such as sentiment, emotion, intention, and opinion (Mai et al., 2025; Poria et al., 2017). Recently, MAC has achieved remarkable progress and become increasingly important for applications such as human-computer interaction, customer service automation, and affective computing systems.

Despite remarkable advances, existing approaches often learn spurious cross-modal correlations from training data rather than genuine causal relationships, harming model generalization when test data distributions differs from training distributions under distribution shifts or noisy modality conditions (Zhuang et al., 2025; Peters et al., 2016). This limitation restricts their practical deployment in real-world scenarios where environmental conditions, speaking styles, lighting conditions, and background contexts constantly vary (Arjovsky et al., 2019). For instance, models might over-rely on the speaker's consistent smile (a spurious visual feature) in training data rather than focusing on semantic content and genuine emotional expressions, leading to performance degradation when tested on different speakers (see Figure 1). Similarly, noisy modalities (e.g., background noise or low-resolution visual frames) further disrupt spurious correlations, exacerbating the generalization gap. To address this limitation, some approaches focus on domain adaptation to learn shared representations across domains (Dai et al., 2020; Zhu et al., 2025; Zhang et al., 2022), while others employ disentanglement strategy to separate different factors in multimodal data (Hazarika et al., 2020; Zhuang et al., 2025). However, these methods lack causal interpretation and cannot guarantee the disentangled/learned features align with causal and spurious components. Recent works incorporate causal inference principles to identify and eliminate spurious correlations (Xu et al., 2025b; Jiang

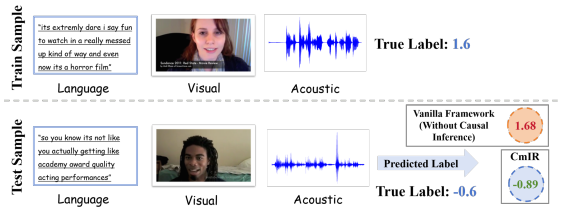


Figure 1: A case study on CMU-MOSI. Vanilla model without causal inference makes incorrect prediction for the test sample where the speaker delivers negative comment while smiling, while CmIR accurately predicts the label based on correct causal relationships.

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45698–45721

et al., 2025; Sun et al., 2022). However, they often lack rigorous theoretical analysis or focus on specific biases (such as speaker bias and modality bias) rather than providing a general framework, which rely on prior knowledge or assumptions about the biases and may not generalize well to unseen environments or other types of distribution shifts.

To resolve these issues, we propose a Causal modality-Invariant Representation (CmIR) learning framework for robust multimodal learning. Drawing on causal inference principles (Arjovsky et al., 2019; Peters et al., 2016), CmIR is theoretically grounded in causal inference and information theory, with a novel feature disentanglement method that separates each modality into two complementary components: (1) **invariant causal representation** (Z_m^{inv}), which carries stable causal relationships with labels across different environments; and (2) **environment-specific spurious representation** (Z_m^{spu}), which captures non-causal, environment-dependent noise and has no causal link to the label. The core insight is that while the distribution of raw features may vary across environments, the causal mechanisms between invariant features and prediction targets remain stable. CmIR is achieved through an elegant objective function that incorporates: (1) an **invariance constraint** to ensure invariant representations maintain stable relationships with labels across environments; (2) a **mutual information constraint** to minimize correlations between invariant and spurious components; and (3) a **reconstruction constraint** to preserve sufficient information from raw inputs. These constraints guarantee that invariant representations satisfy critical properties: environmental independence and causal sufficiency for prediction. We then use invariant modality representations for prediction, ensuring robustness to distribution shifts and noisy modalities. Compared to previous methods (Xu et al., 2025b; Yang et al., 2024), CmIR does not rely on specific bias or assumption. It directly learns invariant representations that are stable across all environments, which is more general and applicable to various distribution shifts.

Our contributions are summarized as follows:

- **Methodological innovation:** We propose a novel framework name CmIR, which, for the first time, systematically disentangles each modality into causal and spurious components in MAC to comprehensively learn causality-sufficient invariant representations.
- **Empirical validation:** Extensive experiments on multiple multimodal tasks show that CmIR achieves state-of-the-art results in both standard and **out-of-distribution** (OOD) benchmarks. Moreover, it exhibits superior robustness under **noisy modality** testing.
- **Theoretical guarantees:** We prove the existence and extractability of invariant representations given multi-environment training data. We also show that predictors based on invariant representations achieve lower worst-case OOD risk than those using raw features.

2 Related Work

2.1 Multimodal Affective Computing

Most works for MAC center on devising fusion techniques to learn discriminative multimodal representations (Zadeh et al., 2018a; Wang et al., 2025; Zadeh et al., 2017) or employing techniques such as information bottleneck to regularize unimodal distributions (Shankar, 2022; Mai et al., 2023c; Luo et al., 2025b). Recently, multimodal large language models have enabled the direct processing of multimodal signals using large pre-trained models, enhancing the interpretation of human affective states (Zhao et al., 2025; Xu et al., 2025a). However, these methods often neglect to improve the generalizability of models for OOD data. To enhance generalization and robustness, some methods focus on domain adaptation to learn shared representations across domains (Dai et al., 2020; Zhu et al., 2025; Zhang et al., 2022), while others employ disentangled learning to separate different factors in multimodal data (Yang et al., 2023; Hazarika et al., 2020; Zhuang et al., 2025; Tsai et al., 2019b). However, these methods lack causal interpretation and cannot guarantee the disentangled/learned features align with causal and spurious components.

2.2 Causal Inference

Causal inference can detect and remove non-causal associations in complex datasets to improve model robustness and generalization (Wang et al., 2022; Niu et al., 2021). Many causality-based methods have been introduced to reduce cross-modal bias in multimodal learning. Researchers employ counterfactual reasoning to refine attention distributions (Huang et al., 2025), apply front-door and back-door adjustments to decouple spurious links between text and vision (Liu et al., 2023), develop counterfactual and debiasing frameworks (Sun

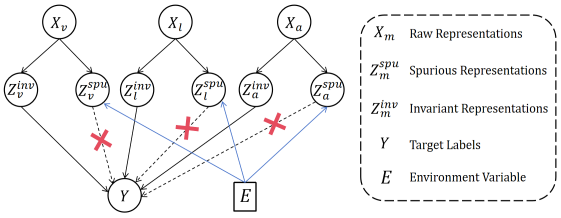


Figure 2: The SCM of CmIR for prediction process. It includes language (l), visual (v), acoustic (a) modalities.

et al., 2022; Huan et al., 2024; Sun et al., 2023), and design causal intervention modules to separate misleading connections between expressive style and semantic content (Xu et al., 2025b). However, most methods are restricted to single or specific modality pairs, or they rely on explicitly annotated bias types that require domain knowledge (Nam et al., 2020). This limitation hinders their broader application to complex multimodal data where biases are often implicit and not predefined. In contrast, we propose a general invariant representation learning framework without requiring predefined bias types. Moreover, Invariant Risk Minimization (Arjovsky et al., 2019) and Invariant Causal Mechanism of CLIP (Song et al., 2025) that aim to learn invariant features are related to our work, but they focus on single modalities or specific modality combinations in particular application scenarios. In contrast, we provide a more general multimodal causal framework with feature disentanglement to understand and learn the properties of invariant features more comprehensively and accurately.

3 Theoretical Analysis

Here we establish a theoretical foundation for our causal approach to multimodal learning. We first establish the existence and extractability of causal invariant modality representations, then prove their advantages in terms of generalization performance.

3.1 Definition of Invariant Representations

We begin by formalizing the causal structure of multimodal learning. Consider M modalities $\{X_m\}_{m=1}^M$ and prediction target Y. Following the structural causal model (SCM) framework (Peters et al., 2016), we assume the data generation process involves an environment variable E that induces distribution shifts, and each modality X_m can be decomposed into an invariant component Z_m^{inv} (containing causal features) and a spurious component Z_m^{spu} (containing environment-specific features). As shown in Figure 2, we assume Y is causally influenced only by invariant components $\{Z_m^{\text{inv}}\}$ and is independent of E and $\{Z_m^{\text{spu}}\}$.

Theorem 1 (Definition of Causal Invariant Modality Representations)**.** *Assume there exists a function class $\Phi_m = \{\phi_m : \mathcal{X}_m \to \mathcal{Z}_m^{inv}\}$ and a distribution distance measure $\mathcal{D}$ such that the following optimization problem has a solution:*

$$\phi_m^* = \arg\min_{\phi_m \in \Phi_m} \max_{e_1, e_2 \in \mathcal{E}} \mathcal{D}(P(Y|\phi_m(X_m), E=e_1), P(Y|\phi_m(X_m), E=e_2))$$

Then $\phi_m^*(X_m)$ constitutes a causal invariant modality representation satisfying:

$$P(Y|\phi_m^*(X_m), E=e_1) = P(Y|\phi_m^*(X_m), E=e_2), \quad \forall e_1, e_2 \in \mathcal{E}$$

Proof. See Section A.1 for the proof. □

Theorem 1 suggests that $\phi_m^*(X_m)$ is a valid invariant modality representation capturing only causal features for robust prediction. If we can learn a representation $\phi(X_m)$ such that the conditional distribution of the label Y given this representation is the same across all environments (i.e., $P(Y|\phi(X_m), E=e)$ does not depend on e), then this representation must capture only the causal features (i.e., causal features exist). Intuitively, causal relationships between features and labels are invariant under changes of the environment. If the prediction rule based on $\phi(X_m)$ remains unchanged when the environment varies, it means that $\phi(X_m)$ does not contain any environment-specific spurious information. In other words, it blocks all backdoor paths from environment E to label Y. Thus, the invariance condition is a signature of causality.

3.2 Extraction of Invariant Representations

While Theorem 1 establishes the definition of invariant representations, practical implementation requires extracting these representations from raw modalities while preserving all relevant information. This motivates our disentanglement approach.

Theorem 2 (Theoretical Guarantee for Extracting Disentangled Representations)**.** *Consider encoder functions $g_m : \mathcal{X}_m \to (\mathcal{Z}_m^{inv}, \mathcal{Z}_m^{spu})$ and decoder functions $r_m : (\mathcal{Z}_m^{inv}, \mathcal{Z}_m^{spu}) \to \mathcal{X}_m$ that optimize the following objective:*

$$\min_{\{g_m, r_m, h\}_{m=1}^M} \mathbb{E}_{e \in \mathcal{E}}[\mathcal{L}_{pred}(Y, h(\{Z_m^{inv}\}_{m=1}^M))] + \lambda_1 \sum_{m=1}^M \mathcal{R}_{inv}^{(m)} + \lambda_2 \sum_{m=1}^M \mathcal{R}_{dec}^{(m)} + \lambda_3 \sum_{m=1}^M \mathcal{R}_{rec}^{(m)}$$

where $\mathcal{L}_{pred}$ *is the task prediction loss,* h *is the prediction head,* $\mathcal{R}_{inv}^{(m)} = \sum_{e_1 \in \mathcal{E}} \sum_{e_2 \in \mathcal{E}} \mathcal{D}(P(Y|Z_m^{inv}, E = e_1), P(Y|Z_m^{inv}, E = e_2))$ *enforces invariance,* $\mathcal{R}_{dec}^{(m)} = I(Z_m^{inv}; Z_m^{spu})$ *minimizes mutual information between invariant and spurious components,* $\mathcal{R}_{rec}^{(m)} = \|X_m - r_m(Z_m^{inv}, Z_m^{spu})\|^2$ *ensures reconstruction capability, and* $\lambda_1, \lambda_2, \lambda_3 > 0$ *are hyperparameters. Assuming the label* Y *is independent of environment* E*, the function classes* $\{g_m, r_m, h\}$ *have sufficient capacity and the data follows the SCM described in Section 3.1, then as* $\lambda_1, \lambda_2, \lambda_3 \to \infty$*, the optimal solution satisfies:*

1. $\lim_{\lambda_3 \to \infty} \mathbb{E}\left[\|X_m - r_m(Z_m^{\text{inv}}, Z_m^{\text{spu}})\|^2\right] = 0$ *(perfect reconstruction is achieved)*
2. $Z_m^{inv} \perp\!\!\!\perp E$ *(invariant component is environment-independent)*
3. $I(Y; Z_m^{spu} | Z_m^{inv}, E) = 0$ *(spurious component contains no additional causal information)*

Proof. See Section A.2 for the proof. □

Theorem 2 suggests that causal invariant representations can be learned using our CmIR. The invariance constraint forces Z^{inv} to have the same predictive relationship with Y across environments, making it environment-independent (capturing only causal features). The reconstruction loss ensures that the pair (Z^{inv}, Z^{spu}) retains all information from the original input, preventing information loss and avoiding degenerate decomposition solutions. The mutual information minimization pushes Z^{inv} and Z^{spu} to be statistically independent, so that Z^{spu} cannot carry any causal information about Y that is already in Z^{inv}. Reconstruction loss and mutual information minimization together force Z^{inv} to contain all causal information in the original input. These three constraints lead to a clean decomposition: Z^{inv} contains only causal factors and encompasses all causal information from the original input, and Z^{spu} contains only environment-specific noise.

3.3 Distributionally Robust Risk Advantage of Invariant Representations

Having established how to learn invariant representations, we now prove their theoretical advantages for worst-case OOD risk under distribution shift.

Theorem 3 (Distributionally Robust Risk Advantage of Invariant Representations)**.** *Let* $\mathcal{H}$ *be a hypothesis class over* $\mathcal{X} \times \mathcal{Y}$*, where* $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M$ *is the* M*-modal feature space and* $\mathcal{Y}$ *is the label space. Let* $\mathcal{E}_{all}$ *denote the set of all possible environments, each corresponding to a distribution* $P^e(x, y)$*. Let* $h_{inv} \in \mathcal{H}$ *be a predictor using invariant representations* $Z^{inv} = \{Z_m^{inv}\}_{m=1}^M$*, and* $h_{raw} \in \mathcal{H}$ *be a predictor using raw multimodal representations* $X = \{X_m\}_{m=1}^M$*. Assume:*

1. ***Invariance Condition****: The invariant representations satisfy* $P^e(Y|Z^{inv}) = P^{e'}(Y|Z^{inv})$ *for all* $e, e' \in \mathcal{E}_{all}$*.*
2. ***Information Sufficiency****: The mutual information between invariant representations and raw features satisfies* $I(Z^{inv}; X) > c$ *for some constant* $c > 0$ *(*Z^{inv} *contains enough information from* X*).*
3. ***Loss Function Regularity****: The loss function* ℓ *is* L*-Lipschitz continuous and bounded.*

Then the worst-case OOD risk satisfies:

$$R^{OOD}(h_{inv}) < R^{OOD}(h_{raw})$$

where $R^{OOD}(h) = \max_{e \in \mathcal{E}_{test}} R^e(h)$ *and* $R^e(h) = \mathbb{E}_{(x,y) \sim P^e}[\ell(h(x), y)]$*.*

Proof. See Section A.3 for the proof. □

Theorem 3 proves that under realistic conditions, predictors based on invariant representations achieve strictly lower worst-case out-of-distribution risk than those using raw features. The intuition is straightforward: raw features contain both causal and spurious parts. The spurious part may change arbitrarily in new environments, causing large errors in the worst-case scenario. In contrast, the invariant representation relies only on the stable causal mechanism, which remains unchanged across environments, thereby guaranteeing more reliable performance even under the most adverse distribution shifts.

4 Algorithm Implementation

Here we elaborate on the implementation of CmIR proposed in Theorem 2. Our objective is to learn the disentangled modality representations, and perform prediction by fusing invariant representations Z_m^{inv}. The overall framework (see Figure 3) consists of unimodal networks that produces raw unimodal features $X_m \in \mathbb{R}^{1 \times d}$ (see Appendix B for unimodal networks), encoders g_m, decoders r_m, and a prediction head (predictor) h. Next, we detail the implementation of each loss component.

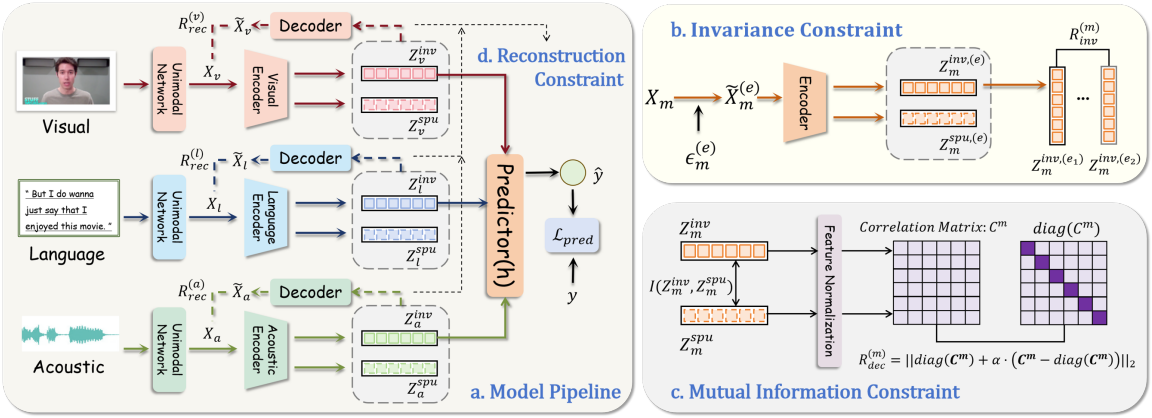


Figure 3: The overall framework of CmIR and the visualization of the proposed constraints.

4.1 Prediction Loss $\mathcal{L}_{\text{pred}}$

The prediction loss ensures that invariant representations effectively predict the target label Y. Given a batch of data, the prediction loss is computed as:

$$Z_m^{\text{inv}},\ Z_m^{\text{spu}} = g_m(X_m) \tag{1}$$

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \ell\left(h\left(\{Z_{m,i}^{\text{inv}}\}_{m=1}^{M}\right), Y_i\right) \tag{2}$$

where g_m is the encoder for modality m, N is the batch size, and ℓ is the corresponding loss function. For classification tasks (e.g., humor detection and sarcasm detection), cross-entropy loss is utilized. For regression tasks (e.g., sentiment analysis), mean squared error (MSE) or mean absolute error (MAE) is used. The predictor is implemented using unimodal feature concatenation and a few multi-layer perception layers (see Figure 7).

4.2 The Invariance Constraint $\mathcal{R}_{\text{inv}}^{(m)}$

The invariance constraint requires that $P(Y|Z_m^{\text{inv}}, E)$ remains invariant across different environments. However, real-world data often lack explicit environment labels E. To address this, we draw inspiration from data augmentation and simulate different **virtual environments** by injecting varying degrees of noise into the raw features X_m. For each sample, we assign a random virtual environment label $e \in \{1, 2, ..., K\}$, where K is the number of environments (a hyperparameter, see Table 7). Then we perform **noise perturbation** via applying an additive noise $\epsilon_m^{(e)} = \alpha^{(e)} \cdot \epsilon_m$ to X_m based on the environment label e:

$$\tilde{X}_m^{(e)} = X_m + \alpha^{(e)} \cdot \epsilon_m, \quad \epsilon_m \sim \mathcal{N}(0, \Sigma_m) \tag{3}$$

where $\alpha^{(e)}$ is an environment-dependent coefficient controlling the noise intensity, and Σ_m is the modality-specific covariance matrix (which can be set as the identity matrix or estimated from the data). The noise coefficient $\alpha^{(e)}$ is different for different environments, which is defined as $\alpha^{(e)} = \alpha^{(1)} * e, e \in \{1, 2, ..., K\}$. $\alpha^{(1)}$ is the noise coefficient for Environment 1, which is a hyperparameter whose values are shown in Table 7. The perturbed feature $\tilde{X}_m^{(e)}$ is fed into the encoder g_m to obtain the invariant representation $Z_m^{\text{inv},(e)}$ for that environment. Finally, the invariance constraint is implemented by minimizing the discrepancy between the conditional distributions $P(Y|Z_m^{\text{inv}}, E)$ across different environments. To realize this, we can adopt a common strategy that enforces consistency in the output distributions of predictor across environments for classification tasks. Specifically, Kullback-Leibler (KL) divergence can be used as the distribution distance measure $\mathcal{D}$:

$$\mathcal{R}_{\text{inv}}^{(m)} = \sum_{e_1 \neq e_2} \text{KL}\left(P(Y|Z_m^{\text{inv},(e_1)}) \| P(Y|Z_m^{\text{inv},(e_2)})\right)$$

where $P(Y|Z_m^{\text{inv},(e)})$ is given by the output of predictor (after Softmax) on the corresponding environment's representation. However, this strategy requires to implement a unimodal predictor for each invariant modality representation which increases the model complexity, and training noise might be introduced if unimodal predictors are not well trained. Moreover, it is hard for regression tasks to calculate the KL-divergence between output distributions. To this end, we adopt a simpler implementation that encourages the learned invariant representations to be identical across different environments, which is a stronger constraint that satisfies the invariance constraint because the out-

put distributions must be the same if input features were the same. We have provided the comparison results of these two variants in Appendix H. Specifically, the invariance constraint is implemented as:

$$\mathcal{R}_{\text{inv}}^{(m)} = \sum_{e_1 \neq e_2} \|Z_m^{\text{inv},(e_1)} - Z_m^{\text{inv},(e_2)}\|_1 \quad (4)$$

Minimizing this term encourages the model to extract features from X_m that are insensitive to noise perturbations (simulating environmental changes).In practice, for each sample in a batch, we can assign an environment label e and generate $\tilde{X}_m^{(e)}$ using Eq. 3. For K environments, we can generate $K+1$ variants of unimodal representations (including the original unimodal representation itself). The invariance loss $\mathcal{R}_{\text{inv}}^{(m)}$ is computed over all $K(K+1)/2$ pairs for each sample in the batch, ensuring strong invariance constraints.

4.3 Mutual Information Constraint $\mathcal{R}_{\text{dec}}^{(m)}$

Theorem 2 requires minimizing the mutual information $I(Z_m^{\text{inv}}; Z_m^{\text{spu}})$ between the invariant representation Z_m^{inv} and the spurious representation Z_m^{spu} to promote their disentanglement and capture independent information. Directly computing mutual information is intractable. We employ a widely used and effective alternative: approximating mutual information minimization by enforcing **orthogonality** (zero linear correlation) between the two representations in the feature space, which is a practical and computationally efficient proxy for minimizing mutual information between Z_m^{inv} and Z_m^{spu}. Orthogonality is a necessary condition for statistical independence, and we augment this constraint with invariance and reconstruction constraints to ensure semantic separation of causal and spurious factors. This proxy is widely used in disentanglement learning for its scalability to large multimodal datasets. Minimizing this term encourages Z_m^{inv} and Z_m^{spu} to learn in orthogonal directions, thereby reducing information redundancy between them.

Specifically, for each batch of training data, the correlation matrix $\boldsymbol{C}^m$ can be calculated as:

$$\boldsymbol{C}^m = Nor(\boldsymbol{Z}_m^{\text{inv}}) Nor(\boldsymbol{Z}_m^{\text{spu}})^\top \quad (5)$$

where $Nor(x) = \frac{x - mean(x)}{std(x)}$ denotes feature normalization, $\boldsymbol{Z}_m^{\text{inv}} \in \mathbb{R}^{N \times d}$ denotes a batch of invariant modality representations, and $\boldsymbol{C}^m \in \mathbb{R}^{N \times N}$ is the correlation matrix for modality m. Then, we enforce orthogonality via the following operation:

$$\mathcal{R}_{\text{dec}}^{(m)} = \|diag(\boldsymbol{C}^m) + \alpha \cdot (\boldsymbol{C}^m - diag(\boldsymbol{C}^m))\|_2 \quad (6)$$

where $diag(\boldsymbol{C}^m)$ denotes the diagonal matrix of $\boldsymbol{C}^m$, and α is a hyperparameter that is between zero and one. In Eq. 6, we use the Frobenius norm of matrix $\|\cdot\|_F$ (standard for matrix regularization). The term α balances the constraint strength between diagonal (same-sample) and off-diagonal (cross-sample) terms in the correlation matrix, which is a hyperparameter that depends on datasets (see Table 7). When α is less than 1, it can down-weight off-diagonal terms to focus on sample-wise orthogonality. In this way, we can enforce a stricter constraint on the invariant and spurious representations from the same sample, and also encourage invariant and spurious representations from different samples to be orthogonal, promoting the statistical independence between two representations.

4.4 Reconstruction Constraint $\mathcal{R}_{\text{rec}}^{(m)}$

The reconstruction loss ensures that the disentangled representations $(Z_m^{\text{inv}}, Z_m^{\text{spu}})$ retain all information from the original input X_m, preventing the loss of crucial content during representation learning.

Firstly, the encoder g_m maps the input X_m to the disentangled representation pair $(Z_m^{\text{inv}}, Z_m^{\text{spu}})$. The decoder r_m then attempts to reconstruct the original input from this pair:

$$\hat{X}_m = r_m(Z_m^{\text{inv}}, Z_m^{\text{spu}}) \quad (7)$$

The reconstruction loss is computed using MSE:

$$\mathcal{R}_{\text{rec}}^{(m)} = \|X_m - \hat{X}_m\|_2^2 \quad (8)$$

The encoder and decoder are implemented as multilayer perceptron networks (see Figure 7).

4.5 Overall Optimization Objective

The complete optimizable objective function is:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \sum_{m=1}^{M} \lambda_1 \mathcal{R}_{\text{inv}}^{(m)} + \lambda_2 \mathcal{R}_{\text{dec}}^{(m)} + \lambda_3 \mathcal{R}_{\text{rec}}^{(m)} \quad (9)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters that balance the importance of each constraint.

5 Experiments

CmIR is evaluated on multiple tasks of MAC, including multimodal sentiment analysis (MSA), multimodal humor detection (MHD) and multimodal sarcasm detection (MSD). The used datasets include CMU-MOSI (Zadeh et al., 2016), CMU-MOSI (OOD) (Sun et al., 2022), CMU-MOSEI

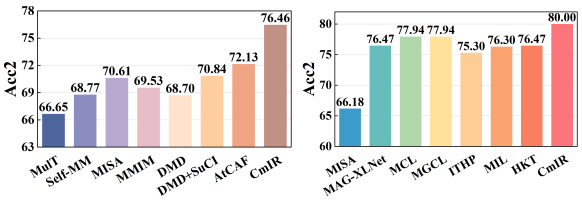
(Zadeh et al., 2018b), CH-SIMS-v2 (Liu et al., 2022), UR-FUNNY (Hasan et al., 2019) and MUStARD (Castro et al., 2019). Due to space limitation, **we introduce experimental settings, baselines, datasets and additional results in Appendix.** Codes are available at: https://github.com/TmacMai/CmIR.

5.1 Performance on the MSA Task

The performance of CmIR on MSA is summarized in Table 1 and Table 2. On CMU-MOSI, CmIR surpasses strong baseline ITHP (Xiao et al., 2024) by more than 2 points in Acc7 and 1 points in Acc2. For CMU-MOSEI, it outperforms GSCon (Shi et al., 2025) and achieves the best scores in Acc2, F1, MAE, and Acc7.Compared with feature disentanglement method FDMER (Yang et al., 2022) and previous backdoor-adjustment work that focuses on specific confounders (Xu et al., 2025b), CmIR demonstrates considerable improvement. Similar superiority is observed on CH-SIMS-v2 (Table 2), where CmIR outperforms all baselines across every metric, including an improvement of 2.5 points in Acc5. Overall, **CmIR establishes state-of-the-art results on MSA across three standard benchmarks**. This strong performance is primarily attributed to CmIR's causal learning strategy, which effectively learns invariant representations across all environments that eliminate general bias and enable a more robust multimodal learning.

5.2 Performance on the MHD and MSD Tasks

To assess the task generalizability of CmIR, we evaluate it on MHD and MSD (classification tasks) using UR-FUNNY and MUStARD datasets. The baselines include MulT (Tsai et al., 2019a), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), HKT (Hasan et al., 2021), DMD (Li et al., 2023), DMD+SuCI (Xu et al., 2025b), AtCAF (Huang et al., 2025), MAG-XLNet (Rahman et al., 2020), MCL (Mai et al., 2023a), MGCL (Mai et al., 2023b), ITHP (Xiao et al., 2024), MISA (Hazarika et al., 2020), and MIL (Zhang et al., 2024), where DMD+SuCI and AtCAF are causality-based methods. As shown in Figure 4, CmIR surpasses the strongest baselines (AtCAF and MGCL) by margins exceeding 4 and 2 points on UR-FUNNY and MUStARD, respectively. Overall, **CmIR achieves competitive performance on both MHD and MSD**, confirming its effectiveness and **strong generalizability to diverse multimodal tasks**.



(a) Results on MHD task (b) Results on MSD task

Figure 4: The results on (a) UR-FUNNY (Hasan et al., 2019) and (b) MUStARD (Castro et al., 2019) datasets.

5.3 Results under OOD scenarios.

The results under OOD scenarios is depicted in Table 3. It is observed that: I) All models degrade when moving from in-distribution to OOD settings, verifying that spurious correlations impede generalization; II) **CmIR delivers significantly stronger OOD performance** than standard multimodal baselines. Its advantage over ITHP (Xiao et al., 2024) grows notably, with Acc2 improvement rising from 1.5 points to 3.5 points, and Acc7 improvement increasing from 2.1 points to 7.2 points, underscoring the efficacy of our causal strategy; III) Compared to recent causality-based methods (CLUE, GEAR, MulDeF), CmIR consistently outperforms them in all metrics, highlighting its robustness in mitigating broad spurious correlations. This is because instead of focusing on specific bias or assumption, CmIR directly learns invariant representations that are stable across all environments, which is more general and applicable to various distribution shifts.

5.4 Discussion on Noisy Modalities

(1) To assess the robustness of CmIR to modality noise, we corrupt all modalities of all training and testing samples with **Gaussian noise** (the noise rate NR is set at 10% -70%). The compared baselines include TMDC (Zhuang et al., 2025), C-MIB (Mai et al., 2023c) and Multimodal Boosting (Mai et al., 2024), which adopt the same training and testing settings as CmIR. Following prior work (Mai et al., 2024), we report Acc2 and MAE. Table 4 shows that **CmIR outperforms competitive baselines across most metrics (particularly in MAE)**, and its **performance advantage becomes even more pronounced as the noise level increases**. This is mainly because CmIR can more accurately identify and extract causal features from noisy inputs and maintain stable predictive ability for labels via the proposed constraints. These results indicate the robustness of CmIR in handling noise data.

(2) To assess the resilience of CmIR to **out-**

Table 1: Comparisons on the CMU-MOSI and CMU-MOSEI datasets. The results labeled with † are obtained from original papers, and other results are obtained from our experiments. The best results are highlighted.

Model	Venue	CMU-MOSI					CMU-MOSEI				
		Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
Self-MM (Yu et al., 2021)	AAAI 21	45.8	84.9	84.8	0.731	0.785	53.0	85.2	85.2	0.540	0.763
ConFEDE (Yang et al., 2023)	ACL 23	43.3	85.1	85.2	0.728	0.784	52.7	85.7	85.6	0.538	0.772
FDMER† (Yang et al., 2022)	ACM MM 22	44.1	84.6	84.7	0.724	0.788	54.1	86.1	85.8	0.536	0.773
SuCI† (Xu et al., 2025b)	AAAI 25	42.2	84.6	84.5	-	-	54.6	85.8	85.7	-	-
C-MIB (Mai et al., 2023c)	TMM 23	47.7	87.8	87.8	0.662	0.835	52.7	86.9	86.8	0.542	0.784
EMOE (Fang et al., 2025)	CVPR 25	45.2	84.8	84.8	0.723	0.790	52.5	85.0	85.0	0.542	0.760
Multimodal Boosting (Mai et al., 2024)	TMM 24	49.1	88.5	88.4	0.634	0.855	54.0	86.5	86.5	0.523	0.779
ITHP (Xiao et al., 2024)	ICLR 24	47.7	88.5	88.5	0.663	**0.856**	52.2	87.1	87.1	0.550	0.792
Diffusion Bridge (Lee et al., 2025)	CVPR 25	47.3	86.9	86.8	0.649	0.839	53.1	87.1	87.0	0.531	**0.800**
GSCon (Shi et al., 2025)	TIP 25	45.7	88.1	88.0	0.696	0.832	50.8	87.4	87.4	0.561	0.750
CmIR	-	**49.8**	**89.6**	**89.5**	**0.616**	0.853	**55.1**	**87.8**	**87.7**	**0.513**	0.793

Table 2: The results on CH-SIMS-v2.

Model	CH-SIMS v2					
	Acc5↑	Acc3↑	Acc2↑	F1↑	MAE↓	Corr↑
MISA (Hazarika et al., 2020)	47.5	68.9	78.2	78.3	0.342	0.671
MAG-BERT (Rahman et al., 2020)	49.2	70.6	77.1	77.1	0.346	0.641
Self-MM (Yu et al., 2021)	53.5	72.7	78.7	78.6	0.315	0.691
MMIM (Han et al., 2021)	50.5	70.4	77.8	77.8	0.339	0.641
AV-MC (Liu et al., 2022)	52.1	73.2	80.6	80.7	0.301	0.721
KuDA (Feng et al., 2024)	53.1	74.3	80.2	80.1	0.289	0.741
DLF (Wang et al., 2025)	47.5	70.0	78.1	77.9	0.346	0.683
Diffusion Bridge (Lee et al., 2025)	52.5	70.7	78.6	79.0	0.323	0.677
CmIR	**56.0**	**75.0**	**81.9**	**82.0**	**0.286**	**0.742**

Table 3: Results on CMU-MOSI (OOD). Following causal-based methods (Sun et al., 2022, 2023), Acc2† and F1† denote results considering neutral samples.

Methods	Acc7	Acc2†	Acc2	F1†	F1
CLUE (Sun et al., 2022)	41.8	78.8	79.9	78.8	79.9
GEAR (Sun et al., 2023)	-	80.5	82.1	80.4	82.1
MulDeF (Huan et al., 2024)	42.9	79.6	81.5	79.7	81.6
Self-MM (Yu et al., 2021)	40.3	76.7	78.1	76.7	78.1
ITHP (Xiao et al., 2024)	41.3	79.5	81.3	79.5	81.3
KAN-MCP (Luo et al., 2025a)	41.8	79.3	81.5	79.1	81.4
CmIR	**48.5**	**83.0**	**84.4**	**83.0**	**84.4**

of-distribution (OOD) noises not encountered in training, we adopt a mixed-noise evaluation strategy: training samples are contaminated with Gaussian noise, while testing samples are perturbed with distinct noise types (Laplace and random erasing). As presented in Table 4, 'CmIR (OOD)' obtains competitive results and significantly surpasses strong baselines. These findings confirm that **CmIR generalizes effectively to unseen noises**, highlighting its promising application potential.

5.5 Ablation Experiments

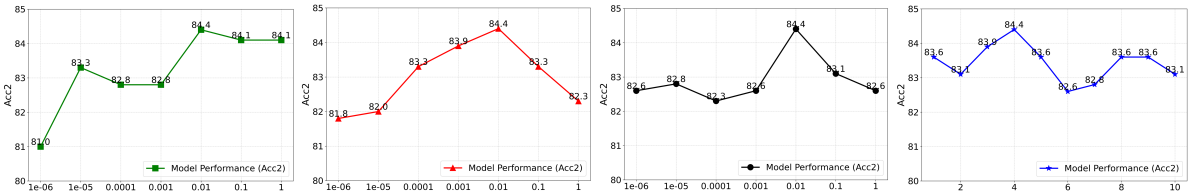
(1) Causal Inference: As presented in Table 5, in the case of 'Vanilla Framework', we directly use the raw modality features for prediction. The model exhibits its sharpest performance decline (over 3 points in Acc2 and Acc7), suggesting the importance of learning causal invariant representations for more robust multimodal prediction and verifying our claim; **(2) Invariance Constraint**: As shown in 'W/O $\mathcal{R}_{\text{inv}}^{(m)}$', removing invariance constraint leads to a noticeable performance drop, because it is the core constraint to learn causal invariant representations. Without it, we cannot ensure that the learned Z_m^{inv} is environment-invariant and the core idea of CmIR cannot be realized; **(3) Mutual Information Constraint**: When $\mathcal{R}_{\text{dec}}^{(m)}$, is removed, the extent of performance decline is similar to that observed when $\mathcal{R}_{\text{inv}}^{(m)}$ is removed, indicating the necessity of minimizing the mutual information between Z_m^{inv} and Z_m^{spu}, and demonstrating the effectiveness of our disentanglement framework. Compared with learning only invariant representations (Song et al., 2025), CmIR simultaneously learns both invariant and spurious representations while minimizing their mutual information, enabling the model to understand and learn the properties of invariant representations more easily and comprehensively; **(4) Reconstruction Constraint**: When $\mathcal{R}_{\text{rec}}^{(m)}$ is removed, the performance also shows a noticeable decline, although the drop is the smallest among all ablations. This occurs because the reconstruction loss ensures the causal and spurious representations fully retain all information from the raw features, thereby preventing information loss and suboptimal disentanglement.

5.6 Hyperparameter Robustness Analysis

We evaluate the effectiveness of hyperparameters on CMU-MOSI (OOD), including the weights for invariance constraint λ_1, mutual information constraint λ_2, reconstruction constraint λ_3, and the number of environments K. As shown in Figure 5 (a), (b), and (c), when the values of weights are small, the performance of CmIR experiences a certain degree of degradation, as the effect of the constraints is not fully utilized. Among them, the performance drop is most pronounced when the weight of invariance constraint decreases, highlighting its importance. Conversely, when the values of λ_2 and λ_3 are too large, the performance also declines, likely because mutual information

Table 4: Discussion on noisy modalities on CMU-MOSI and CMU-MOSEI. MuBo denotes Multimodal Boosting.

	NR	Gaussian Noise				OOD Noises (Laplace Noise and Random Erasing Noise)			
		TMDC	**C-MIB**	**MuBo**	**CmIR**	**TMDC**	**C-MIB**	**MuBo**	**CmIR**
		Acc2/MAE	Acc2/MAE	Acc2/MAE	Acc2/MAE	Acc2/MAE	Acc2/MAE	Acc2/MAE	Acc2/MAE
MOSI	0.1	87.4 / 0.748	87.8 / 0.670	86.7 / 0.678	**88.1 / 0.615**	87.2 / 0.769	**87.8** / 0.666	87.4 / 0.639	**87.8 / 0.638**
	0.2	86.6 / 0.741	**87.5** / 0.726	86.1 / 0.738	87.4 / **0.621**	86.7 / 0.861	87.5 / 0.689	87.3 / 0.681	**88.2 / 0.644**
	0.3	86.9 / 0.733	86.4 / 0.912	86.4 / 0.785	**87.3 / 0.650**	86.8 / 0.741	85.1 / 1.019	**87.1** / 0.710	86.9 / **0.678**
	0.4	85.5 / 0.792	83.2 / 1.366	85.5 / 0.841	**86.4 / 0.669**	85.2 / 0.772	85.6 / 1.303	85.3 / 0.900	**85.8 / 0.663**
	0.5	85.0 / 0.912	84.9 / 1.660	86.1 / 1.172	**86.9 / 0.685**	84.6 / 0.867	83.9 / 1.900	**86.9** / 1.114	85.6 / **0.715**
	0.6	84.8 / 1.181	80.8 / 2.595	82.0 / 1.355	**87.1 / 0.680**	83.1 / 0.884	86.5 / 2.272	85.0 / 1.060	**87.0 / 0.689**
	0.7	84.0 / 1.277	82.1 / 3.146	84.4 / 1.750	**86.1 / 0.740**	82.0 / 0.966	84.0 / 3.516	84.7 / 1.731	**85.6 / 0.771**
	Avg	85.7 / 0.912	84.7 / 1.582	85.3 / 1.046	**87.0 / 0.666**	85.1 / 0.837	85.8 / 1.624	86.2 / 0.976	**86.7 / 0.685**
MOSEI	0.1	86.6 / 0.618	86.1 / 0.545	86.4 / 0.544	**87.5 / 0.528**	86.4 / 0.613	86.9 / 0.564	86.1 / 0.561	**87.1 / 0.523**
	0.2	86.2 / 0.593	84.5 / 0.582	86.6 / 0.557	**87.2 / 0.522**	85.5 / 0.631	85.1 / 0.594	85.2 / 0.585	**86.6 / 0.523**
	0.3	85.3 / 0.603	85.6 / 0.622	85.5 / 0.623	**86.8 / 0.531**	85.0 / 0.667	85.9 / 0.665	85.4 / 0.610	**86.5 / 0.528**
	0.4	84.4 / 0.596	84.4 / 0.703	85.3 / 0.682	**86.6 / 0.535**	84.9 / 0.682	84.8 / 0.767	84.4 / 0.714	**86.7 / 0.535**
	0.5	84.6 / 0.592	83.7 / 0.875	84.1 / 0.724	**85.9 / 0.558**	84.3 / 0.751	73.1 / 0.768	84.4 / 0.763	**86.3 / 0.563**
	0.6	83.8 / 0.602	82.4 / 1.054	85.4 / 0.924	**85.8 / 0.548**	82.7 / 0.829	82.0 / 0.856	84.1 / 0.795	**86.3 / 0.644**
	0.7	83.4 / **0.623**	80.5 / 1.404	80.3 / 1.125	**86.3** / 0.671	81.7 / 0.888	**86.5** / 2.366	85.0 / 1.158	85.4 / **0.663**
	Avg	84.9 / 0.604	83.9 / 0.826	84.8 / 0.740	**86.6 / 0.556**	84.4 / 0.723	83.5 / 0.940	84.9 / 0.741	**86.4 / 0.568**

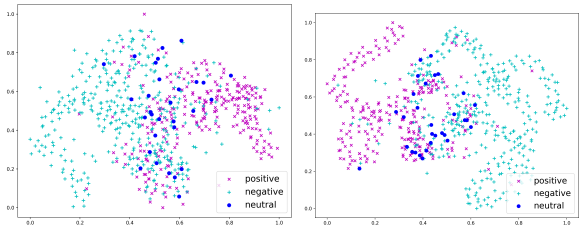


(a) Weight λ_1 (b) Weight λ_2 (c) Weight λ_3 (d) Environment Number K

Figure 5: Acc2 of CmIR w.r.t the change of constraint weights and the number of environments.

Table 5: Ablation experiments on CMU-MOSI.

Model	**Acc7↑**	**Acc2↑**	**MAE↓**
Vanilla Framework	46.3	85.8	0.681
W/O $\mathcal{R}_{\text{inv}}^{(m)}$	45.7	86.7	0.652
W/O $\mathcal{R}_{\text{dec}}^{(m)}$	45.7	86.9	0.671
W/O $\mathcal{R}_{\text{rec}}^{(m)}$	47.7	88.1	0.623
CmIR	**49.8**	**89.6**	**0.616**



(a) Regular Features (b) Invariant Features

Figure 6: T-SNE visualization of language features without/with causal inference learning.

and reconstruction constraints dominate the learning process, preventing sufficient attention to the invariance constraint and the prediction loss. Moreover, as shown in Figure 5 (d), the performance remains relatively stable as the value of K varies, indicating the robustness of CmIR. Overall, **when hyperparameters are varied across a wide range, CmIR's performance consistently maintains a good level** (Acc2 $\geq$ 81%), which to some extent demonstrates the stability of CmIR.

5.7 Visualization of Invariant Representations

To demonstrate that CmIR indeed learns invariant representations that maintain stable predictive power for labels, we intervene with OOD noise on test samples and visualize the extracted causal language representations, alongside visualizing the language representations obtained without causal inference training. As shown in Figure 6, the causal representations learned by CmIR for different classes are well-separated in the feature space, with neutral samples concentrated between the positive and negative ones. In contrast, regular representations learned without CmIR exhibit substantial overlap, and neutral samples appear more scattered. This indicates that **even under noisy conditions, the causal representations learned by CmIR more accurately reflect label information**.

6 Conclusion

We propose CmIR for robust multimodal learning. By disentangling each modality into invariant causal and spurious components, CmIR learns stable and causally-aware representations. We provide theoretical guarantees for CmIR, and demonstrate state-of-the-art results in multiple tasks. CmIR excels under distribution shifts and noisy-modality conditions, highlighting its practical robustness.

Limitations

While CmIR demonstrates strong performance and robustness, our work has certain limitations. First, the environmental simulation via feature perturbation, while effective, may not fully capture the complexity of real-world distribution shifts. Future work could explore more sophisticated environment generation strategies or incorporate real-world multi-environment datasets. Second, the mutual information minimization constraint implemented via feature orthogonality is an approximation. More precise mutual information estimation techniques could be integrated for potentially better disentanglement, albeit at increased computational cost. These limitations, however, do not undermine the core theoretical contributions or the empirical effectiveness of the proposed framework.

7 Acknowledgment

This work is supported by the Guangdong Philosophy and Social Sciences Planning Project (No. GD26YJY34).

References

Emile HL Aarts and 1 others. 1987. *Simulated annealing: Theory and applications*. Reidel.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. In *International Conference on Learning Representations*.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Yong Dai, Jian Liu, Xiancong Ren, and Zenglin Xu. 2020. Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7618–7625.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep: A collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Florian Eyben. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pages 1459–1462.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14314–14324.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. 2024. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14755–14766.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12972–12980.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056.

Devamanyu Hazarika, R. Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *ACM MM*, pages 1122–1131.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. 2024. Muldef: A model-agnostic debiasing framework for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. 2025. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725.

Menghua Jiang, Yuxia Lin, Baoliang Chen, Haifeng Hu, Yuncheng Jiang, and Sijie Mai. 2025. Disentangling bias by modeling intra-and inter-modal causal attention for multimodal sentiment analysis. *arXiv preprint arXiv:2508.04999*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. 2025. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4050–4059.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

Yang Liu, Guanbin Li, and Liang Lin. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. 2025a. Towards explainable fusion and balanced learning in multimodal sentiment analysis. In *ACM MM*, pages 1997–2006.

Yuanyi Luo, Wei Liu, Qiang Sun, Sirui Li, Jichunyang Li, Rui Wu, and Xianglong Tang. 2025b. Triagedmsa: Triaging sentimental disagreement in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. 2024. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, 26:3018–3033.

Sijie Mai, Ya Sun, Ying Zeng, and Haifeng Hu. 2023a. Excavating multimodal correlation for representation learning. *Information Fusion*, 91:542–555.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2023b. Learning from the global view: Supervised contrastive learning of multimodal representation. *Information Fusion*, 100:101920.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2023c. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2025. Learning by comparing: Boosting multimodal affective computing through ordinal learning. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*, pages 2120–2134, New York, NY, USA. Association for Computing Machinery.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2023d. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Debiasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and E. Hoque. 2020. Integrating multimodal information in large pretrained transformers. *ACL*, 2020:2359–2369.

Shiv Shankar. 2022. Multimodal fusion via cortical network inspired losses. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1167–1178.

QingHongYa Shi, Mang Ye, Wenke Huang, Bo Du, and Xiaofen Zong. 2025. Gradient and structure consistency in multimodal emotion recognition. *IEEE Transactions on Image Processing*.

Zeen Song, Siyu Zhao, Xingyu Zhang, Jiangmeng Li, Changwen Zheng, and Wenwen Qiang. 2025. Learning invariant causal mechanism from vision-language models. In *Forty-second International Conference on Machine Learning*.

Teng Sun, Juntong Ni, Wenjie Wang, Liqiang Jing, Yinwei Wei, and Liqiang Nie. 2023. General debiasing for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5861–5869.

Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 15–23.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569.

Yao Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. In *ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 3562–3571.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. 2024. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Zhi Xu, Dingkang Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. 2025b. Debiased multimodal understanding for human language sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14450–14458.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pages 464–481. Springer.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2024. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1114–1125.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *AAAI*, pages 5634–5641.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. 2024. Learning multitask commonness and uniqueness for multimodal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*, 5(3):1349–1361.

Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan. 2022. Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE transactions on neural networks and learning systems*, 34(10):7956–7966.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, and 1 others. 2025. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Multimodal invariant sentiment representation learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14743–14755.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. 2025. Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. *arXiv preprint arXiv:2511.10325*.

A Detailed Theoretical Analysis

Here we establish a detailed theoretical foundation for our causal approach to multimodal learning. We first establish the existence/definition and extractability of causal invariant modality representations, then prove their advantages in terms of generalization performance.

A.1 Definition of Invariant Representations

Theorem 1 (Definition of Causal Invariant Modality Representations) Assume there exists a function class $\Phi_m = \{\phi_m : \mathcal{X}_m \to \mathcal{Z}_m^{\text{inv}}\}$ and a distribution distance measure $\mathcal{D}$ (e.g., KL divergence or Wasserstein distance) such that the following optimization problem has a solution:

$$\phi_m^* = \arg\min_{\phi_m \in \Phi_m} \max_{e_1, e_2 \in \mathcal{E}} \mathcal{D}(P(Y|\phi_m(X_m), E=e_1), P(Y|\phi_m(X_m), E=e_2))$$

Then $\phi_m^*(X_m)$ constitutes a causal invariant modality representation satisfying:

$$P(Y|\phi_m^*(X_m), E=e_1) = P(Y|\phi_m^*(X_m), E=e_2), \quad \forall e_1, e_2 \in \mathcal{E}$$

Proof. The proof follows from the invariance principle in causal inference (Peters et al., 2016). We proceed in three steps:

Step 1 (Necessity of Invariance): If a feature Z contains causal information about Y, then the conditional distribution $P(Y|Z)$ should remain invariant across different environments (Arjovsky et al., 2019). This is because causal mechanisms are stable under interventions on non-descendant variables in the causal graph.

Step 2 (Sufficiency of the Optimization Objective): The optimization objective directly minimizes the maximum discrepancy of $P(Y|\phi_m(X_m), E)$ across environments. By the properties of the distribution distance measure $\mathcal{D}$, $\mathcal{D}(P_1, P_2) = 0$ if and only if $P_1 = P_2$. Therefore, the optimal solution ϕ_m^* must satisfy:

$$\mathcal{D}(P(Y|\phi_m^*(X_m), E=e_1), P(Y|\phi_m^*(X_m), E=e_2)) = 0$$

for all environment pairs $e_1, e_2 \in \mathcal{E}$, which implies:

$$P(Y|\phi_m^*(X_m), E=e_1) = P(Y|\phi_m^*(X_m), E=e_2)$$

Step 3 (Causal Interpretation): The condition $P(Y|\phi_m^*(X_m), E=e_1) = P(Y|\phi_m^*(X_m), E=e_2)$ for all e_1, e_2 implies that $\phi_m^*(X_m)$ blocks all backdoor paths from E to Y that pass through modality m. By the backdoor criterion (Peters et al., 2016), this means $\phi_m^*(X_m)$ contains only causal features from Z_m^{inv} and excludes spurious features from Z_m^{spu}, as the latter would create environment-dependent associations with Y.

This completes the proof that $\phi_m^*(X_m)$ is a valid causal invariant modality representation capturing only causal features. □

A.2 Extractability of Invariant Representations

While Theorem 1 establishes the definition of invariant representations, practical implementation requires extracting these representations from raw modalities while preserving all relevant information. This motivates our disentanglement approach.

Theorem 2 (Theoretical Guarantee for Disentangled Representations) Consider encoder functions $g_m : \mathcal{X}_m \to (\mathcal{Z}_m^{\text{inv}}, \mathcal{Z}_m^{\text{spu}})$ and decoder functions $r_m : (\mathcal{Z}_m^{\text{inv}}, \mathcal{Z}_m^{\text{spu}}) \to \mathcal{X}_m$ that optimize the following objective:

$$\min_{\{g_m, r_m, h\}_{m=1}^M} \mathbb{E}_{e \in \mathcal{E}}[\mathcal{L}_{\text{pred}}(Y, h(\{Z_m^{\text{inv}}\}_{m=1}^M))] + \lambda_1 \sum_{m=1}^M \mathcal{R}_{\text{inv}}^{(m)} + \lambda_2 \sum_{m=1}^M \mathcal{R}_{\text{dec}}^{(m)} + \lambda_3 \sum_{m=1}^M \mathcal{R}_{\text{rec}}^{(m)}$$

where:

- $\mathcal{L}_{\text{pred}}$ is the prediction loss and h is the prediction head
- $\mathcal{R}_{\text{inv}}^{(m)} = \sum_{e_1 \in \mathcal{E}} \sum_{e_2 \in \mathcal{E}} \mathcal{D}(P(Y|Z_m^{\text{inv}}, E = e_1), P(Y|Z_m^{\text{inv}}, E = e_2))$ enforces invariance

- $\mathcal{R}_{\text{dec}}^{(m)} = I(Z_m^{\text{inv}}; Z_m^{\text{spu}})$ minimizes mutual information between invariant and spurious components
- $\mathcal{R}_{\text{rec}}^{(m)} = \|X_m - r_m(Z_m^{\text{inv}}, Z_m^{\text{spu}})\|^2$ ensures reconstruction capability
- $\lambda_1, \lambda_2, \lambda_3 > 0$ are hyperparameters

Assuming the label Y is independent of environment E, the function classes $\{g_m, r_m, h\}$ have sufficient capacity and the data follows the SCM described in Section 3.1, then as $\lambda_1, \lambda_2, \lambda_3 \to \infty$, the optimal solution satisfies:

1. $\lim_{\lambda_3 \to \infty} \mathbb{E}\big[\|X_m - r_m(Z_m^{\text{inv}}, Z_m^{\text{spu}})\|^2\big] = 0$ (perfect reconstruction is achieved)
2. $Z_m^{\text{inv}} \perp\!\!\!\perp E$ (invariant component is environment-independent)
3. $I(Y; Z_m^{\text{spu}} | Z_m^{\text{inv}}, E) = 0$ (spurious component contains no additional causal information)

Proof. We prove the three claims in order. The limit $\lambda_i \to \infty$ is understood in the sense of **tightening constraints**: as λ_i grows, the corresponding regularization term must vanish to keep the loss finite, provided the optimal loss remains bounded. We assume the feasible set (where all terms are finite) is non-empty, which is reasonable given sufficient model capacity.

Part 1 (Perfect Reconstruction): The term $\lambda_3 \mathcal{R}_{\text{rec}}^{(m)}$ with $\lambda_3 \to \infty$ forces the reconstruction error to zero. By the properties of the squared $L2$ norm, $\lim_{\lambda_3 \to \infty} \mathbb{E}\big[\|X_m - r_m(Z_m^{\text{inv}}, Z_m^{\text{spu}})\|^2\big] = 0$ if and only if $X_m = r_m(Z_m^{\text{inv}}, Z_m^{\text{spu}})$ almost surely. Thus, in the limit, the decoder can reconstruct the input with arbitrarily small error. Note that **exact** zero error may be unattainable for finite-dimensional representations, but the limiting statement suffices for theoretical analysis; in practice, taking λ_3 sufficiently large yields negligible reconstruction error. This ensures the disentangled representations $(Z_m^{\text{inv}}, Z_m^{\text{spu}})$ preserve all information in the original modality X_m.

The reconstruction constraint $\mathcal{R}_{\text{rec}}^{(m)}$ is crucial for preventing degenerate solutions. It ensures that the disentangled representations form a sufficient statistic for X_m, preserving all information while separating causal from non-causal components. This is essential for maintaining performance in the source environments while improving generalization to new environments

Part 2 (Environment Independence of Invariant Component): The term $\lambda_1 \mathcal{R}_{\text{inv}}^{(m)}$ with $\lambda_1 \to \infty$ forces $\mathcal{D}(P(Y|Z_m^{\text{inv}}, E = e_1), P(Y|Z_m^{\text{inv}}, E = e_2)) = 0$ for all e_1, e_2. By Theorem 1, this implies $P(Y|Z_m^{\text{inv}}, E = e_1) = P(Y|Z_m^{\text{inv}}, E = e_2)$ for all e_1, e_2.

Now, assume for contradiction that Z_m^{inv} is not independent of E. Then there exist values z^{inv}, e_1, e_2 such that $P(Z_m^{\text{inv}} = z^{\text{inv}}|E = e_1) \neq P(Z_m^{\text{inv}} = z^{\text{inv}}|E = e_2)$. By the law of total probability:

$$P(Y|E = e_i) = \int P(Y|Z_m^{\text{inv}} = z^{\text{inv}}) \times P(Z_m^{\text{inv}} = z^{\text{inv}}|E = e_i) dz^{\text{inv}}$$

Since $P(Y|Z_m^{\text{inv}}) = P(Y|Z_m^{\text{inv}})$ but $P(Z_m^{\text{inv}}|E = e_1) \neq P(Z_m^{\text{inv}}|E = e_2)$, we must have $P(Y|E = e_1) \neq P(Y|E = e_2)$. However, in our SCM, Y is causally independent of E given the invariant features $\{X_m^{\text{inv}}\}$, and consequently given $\{Z_m^{\text{inv}}\}$. This contradiction implies $Z_m^{\text{inv}} \perp\!\!\!\perp E$.

Part 3 (No additional causal information in Z_m^{spu}): We prove $I(Y; Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E) = 0$ by contradiction, using the results from Part 1 and Part 2, the mutual information constraint, and the construction of virtual environments.

Assume, for contradiction, that

$$I := I(Y; Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E) > 0$$

From Part 2, the invariance constraint gives $P(Y \mid Z_m^{\text{inv}}, E) = P(Y \mid Z_m^{\text{inv}})$; hence

$$I = H(Y \mid Z_m^{\text{inv}}) - H(Y \mid Z_m^{\text{inv}}, Z_m^{\text{spu}}, E) \quad (1)$$

The inequality $I > 0$ implies

$$H(Y \mid Z_m^{\text{inv}}, Z_m^{\text{spu}}, E) < H(Y \mid Z_m^{\text{inv}}). \quad (2)$$

Thus, conditioning on (Z_m^{spu}, E) strictly reduces the entropy of Y compared to conditioning on Z_m^{inv} alone. In particular, there exists a measurable set of positive measure on which the conditional distribution $P(Y \mid Z_m^{\text{inv}}, Z_m^{\text{spu}})$ depends on Z_m^{spu} in a non-degenerate way.

By Part 1 (perfect reconstruction in the limit), the pair $(Z_m^{\text{inv}}, Z_m^{\text{spu}})$ determines X_m almost surely. Since Y is a function of the multimodal input, we can write:

$$Y = \Psi\big(Z_m^{\text{inv}}, Z_m^{\text{spu}}, \eta\big)$$

where η is an independent noise term capturing irreducible uncertainty. The dependence on Z_m^{spu} is essential because of (2).

Now consider two different environments e_1 and e_2, because the encoder g_m is deterministic and the same for all environments, the conditional distribution of Z_m^{spu} given Z_m^{inv} and E changes with e. Formally, the map $e \mapsto P(Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E = e)$ is injective; i.e., for $e_1 \neq e_2$ we have

$$P(Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E = e_1) \neq P(Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E = e_2)$$

in the sense that the two conditional distributions are not equal almost everywhere.

Using the law of total probability, for any environment e,

$$P(Y \mid Z_m^{\text{inv}}, E = e) = \int P(Y \mid Z_m^{\text{inv}}, Z_m^{\text{spu}})\, dP(Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E = e) \tag{3}$$

Because the environments can vary significantly, the family of mixing distributions $\{P(Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E = e)\}_{e \in \mathcal{E}}$ is rich enough to distinguish different integrands. In particular, if the integrand $P(Y \mid Z_m^{\text{inv}}, Z_m^{\text{spu}})$ is not constant in Z_m^{spu} (which follows from $I > 0$), then the value of the integral in (3) changes continuously with e when $\alpha^{(e)}$ varies. Hence for $e_1 \neq e_2$, the two integrals cannot be equal. Consequently,

$$P(Y \mid Z_m^{\text{inv}}, E = e_1) \neq P(Y \mid Z_m^{\text{inv}}, E = e_2)$$

which contradicts Part 2 where we established that $P(Y \mid Z_m^{\text{inv}}, E = e)$ is constant across all environments. Therefore, our assumption $I > 0$ is false, and we must have

$$I(Y; Z_m^{\text{spu}} \mid Z_m^{\text{inv}}, E) = 0$$

Together, these three parts prove that the optimal solution satisfies all three claimed properties. □

Remark on finite λ. The limit $\lambda_i \to \infty$ is an idealization. In practice, taking λ_i sufficiently large (but finite) yields approximations where each regularization term is bounded by a small tolerance ϵ, and the conclusions hold up to ϵ errors. This is standard in constrained optimization and does not affect the practical validity of the theorem.

A.3 Distributionally Robust Risk Advantage of Invariant Representations

Having established how to obtain causal invariant modality representations, we now prove their theoretical advantages for worst-case out-of-distribution risk under distribution shift.

Theorem 3 (Distributionally Robust Risk Advantage of Invariant Representations) Let $\mathcal{H}$ be a hypothesis class over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M$ is the M-modal feature space and $\mathcal{Y}$ is the label space. Let $\mathcal{E}_{\text{all}}$ denote the set of all possible environments, each corresponding to a distribution $P^e(x, y)$. Let $h_{\text{inv}} \in \mathcal{H}$ be a predictor using invariant representations $Z^{inv} = \{Z_m^{\text{inv}}\}_{m=1}^M$, and $h_{\text{raw}} \in \mathcal{H}$ be a predictor using raw multimodal representations $X = \{X_m\}_{m=1}^M$. Assume:

1. **Invariance Condition**: The invariant representations satisfy $P^e(Y|Z^{\text{inv}}) = P^{e'}(Y|Z^{\text{inv}})$ for all $e, e' \in \mathcal{E}_{\text{all}}$.

2. **Information Sufficiency**: The mutual information between invariant representations and raw features satisfies $I(Z^{\text{inv}}; X) > c$ for some constant $c > 0$ (Z^{inv} contains enough information related to X).

3. **Loss Function Regularity**: The loss function ℓ is L-Lipschitz continuous and bounded.

Then the worst-case out-of-distribution risk satisfies:

$$R^{\text{OOD}}(h_{\text{inv}}) < R^{\text{OOD}}(h_{\text{raw}})$$

where $R^{\text{OOD}}(h) = \max_{e \in \mathcal{E}_{\text{test}}} R^e(h)$ and $R^e(h) = \mathbb{E}_{(x,y)\sim P^e}[\ell(h(x), y)]$.

Proof. We prove the theorem through four key steps, with additional explanations regarding the validity of assumptions and practical implications.

Step 1 (Information Retention Property):

The information sufficiency assumption $I(Z^{\text{inv}}; X) > c$ is typically reasonable in practice because invariant features often capture fundamental aspects of data that remain stable across environments. For instance, in vision-language tasks, semantic content tends to remain consistent even when visual appearance changes. Formally, by the mutual information chain rule:

$$\begin{aligned} I(Y; X) &= I(Y; Z^{\text{inv}}, Z^{\text{spu}}) = I(Y; Z^{\text{inv}}) \\ &\quad + I(Y; Z^{\text{spu}}|Z^{\text{inv}}) - I(Z^{\text{inv}}; Z^{\text{spu}}|Y) \end{aligned}$$

Since $I(Y; Z^{\text{spu}}|Z^{\text{inv}}) \leq H(Z^{\text{spu}}|Z^{\text{inv}})$ and $I(Z^{\text{inv}}; Z^{\text{spu}}|Y) \geq 0$, we have:

$$\begin{aligned} I(Y; X) &\leq I(Y; Z^{\text{inv}}) + H(Z^{\text{spu}}|Z^{\text{inv}}) \\ &= I(Y; Z^{\text{inv}}) + H(X|Z^{\text{inv}}) \\ &\leq I(Y; Z^{\text{inv}}) + H(X) - I(X; Z^{\text{inv}}) \end{aligned}$$

Rearranging terms:

$$I(Y; Z^{\text{inv}}) \geq I(Y; X) - (H(X) - I(X; Z^{\text{inv}}))$$

Setting $\epsilon(c) = H(X) - I(X; Z^{\text{inv}})$, when $I(X; Z^{\text{inv}}) > c$, we have $\epsilon(c) < H(X) - c$, ensuring that $I(Y; Z^{\text{inv}})$ approaches $I(Y; X)$ as c increases. This indicates that the invariant representations Z^{inv} retain most of the predictive information about Y that is present in the raw features X.

Moreover, regrading assumption 3, in real-world multimodal learning, the Lipschitz continuity assumption is widely applicable. For classification tasks using cross-entropy loss, when input features are normalized (as is common practice), the loss becomes Lipschitz continuous. Similarly, mean squared error loss for regression tasks is inherently Lipschitz continuous. This regularity ensures stable optimization and meaningful generalization bounds. Since ℓ is assumed to be L-Lipschitz continuous, the risk gap is bounded (Song et al., 2025):

$$R^e(h^*_{\text{inv}}) \leq R^e(h^*_{\text{raw}}) + L \cdot \epsilon(c)$$

where h^*_{inv} and h^*_{raw} are the optimal predictors using invariant and raw representations respectively. This shows that when c is sufficiently large, Z^{inv} almost completely preserves the information needed for prediction.

Step 2 (Environment Invariance Property): The invariance condition is realistic in many applications where causal factors remain stable despite environmental changes. For example, in object recognition, shape and category remain invariant while lighting, pose, and background may vary. This property ensures that h_{inv} exhibits consistent performance across environments:

$$R^e(h_{\text{inv}}) = R^{e'}(h_{\text{inv}}) = C, \quad \forall e, e' \in \mathcal{E}_{\text{all}}$$

where C is a constant. Consequently:

$$R^{\text{OOD}}(h_{\text{inv}}) = \max_{e \in \mathcal{E}_{\text{test}}} R^e(h_{\text{inv}}) = C$$

This consistency is a key advantage of invariant representations in distribution shift scenarios.

Step 3 (Risk Variability of Raw Representations): For predictor h_{raw}, the risk variability $\sigma = \max_{e,e' \in \mathcal{E}_{\text{test}}} |R^e(h_{\text{raw}}) - R^{e'}(h_{\text{raw}})|$ is typically large in real-world applications where distribution shifts are significant. This variability stems from the model's reliance on features that change with environment.

Let $R_{\max} = \max_{e \in \mathcal{E}_{\text{test}}} R^e(h_{\text{raw}})$ and $R_{\min} = \min_{e \in \mathcal{E}_{\text{test}}} R^e(h_{\text{raw}})$. The average risk $\bar{R} = \mathbb{E}_{e \in \mathcal{E}_{\text{test}}}[R^e(h_{\text{raw}})]$ satisfies:

$$\begin{aligned} R_{\max} - \bar{R} &= \frac{1}{|\mathcal{E}_{\text{test}}|} \sum_{e \in \mathcal{E}_{\text{test}}} (R_{\max} - R^e(h_{\text{raw}})) \\ &\geq \frac{R_{\max} - R_{\min}}{|\mathcal{E}_{\text{test}}|} = \frac{\sigma}{|\mathcal{E}_{\text{test}}|} \end{aligned}$$

which implies the existence of environment e^* such that:

$$R^{e^*}(h_{\text{raw}}) \geq \bar{R} + \frac{\sigma}{|\mathcal{E}_{\text{test}}|}$$

In practice, σ is often substantial when dealing with significant domain shifts, making this inequality practically relevant.

Step 4 (Worst-Case Risk Comparison): This step reveals why invariant predictors typically outperform raw predictors in worst-case scenarios. The generalization errors δ_{raw} and δ_{inv} represent the gap between training and testing performance.

Notably, δ_{inv} is typically smaller than δ_{raw} because invariant features generalize better across environments. In contrast, σ is often large in real-world scenarios with significant distribution shifts. For example, in cross-domain sentiment analysis, performance can vary dramatically between domains (e.g., movie reviews vs. product reviews), resulting in large σ values.

From Steps 1 and 2:

$$R^{\text{OOD}}(h_{\text{inv}}) = C \leq R^e(h^*_{\text{raw}}) + L \cdot \epsilon(c), \quad \forall e \in \mathcal{E}_{\text{test}}$$

By empirical risk minimization theory, standard generalization bounds apply:

$$R^e(h_{\text{inv}}) \leq R^e(h^*_{\text{inv}}) + \delta_{\text{inv}}, \quad \forall e \in \mathcal{E}_{\text{test}}$$

where δ_{inv} can be very small by realization using an expressive and suitable neutral network for the predictor, which decreases with the number of training samples and depend on model complexity. Moreover, as h^*_{raw} is the optimal predictor for raw features, we have:

$$R^e(h_{\text{raw}}) \geq R^e(h^*_{\text{raw}}), \quad \forall e \in \mathcal{E}_{\text{test}}$$

From Step 3:

$$\begin{aligned} R^{\text{OOD}}(h_{\text{raw}}) &\geq R^{e^*}(h_{\text{raw}}) \\ &\geq \bar{R} + \frac{\sigma}{|\mathcal{E}_{\text{test}}|} \\ &\geq \mathbb{E}_{e \in \mathcal{E}_{\text{test}}}[R^e(h^*_{\text{raw}})] + \frac{\sigma}{|\mathcal{E}_{\text{test}}|} \end{aligned}$$

Similarly:

$$
\begin{aligned}
R^{\text{OOD}}(h_{\text{inv}}) &= C \\
&= \mathbb{E}_{e\in\mathcal{E}_{\text{test}}}[R^e(h_{\text{inv}})] \\
&\leq \mathbb{E}_{e\in\mathcal{E}_{\text{test}}}[R^e(h^*_{\text{raw}})] + L\cdot\epsilon(c) + \delta_{\text{inv}}
\end{aligned}
$$

Combining these results:

$$
\begin{aligned}
&R^{\text{OOD}}(h_{\text{raw}}) - R^{\text{OOD}}(h_{\text{inv}}) \\
&\geq \left(\mathbb{E}_{e\in\mathcal{E}_{\text{test}}}[R^e(h^*_{\text{raw}})] + \frac{\sigma}{|\mathcal{E}_{\text{test}}|}\right) \\
&\quad - (\mathbb{E}_{e\in\mathcal{E}_{\text{test}}}[R^e(h^*_{\text{raw}})] + L\cdot\epsilon(c) + \delta_{\text{inv}}) \\
&= -L\cdot\epsilon(c) - \delta_{\text{inv}} + \frac{\sigma}{|\mathcal{E}_{\text{test}}|}
\end{aligned}
$$

The inequality $R^{\text{OOD}}(h_{\text{raw}}) > R^{\text{OOD}}(h_{\text{inv}})$ holds when:

$$
\sigma > |\mathcal{E}_{\text{test}}|(L\cdot\epsilon(c) + \delta_{\text{inv}})
$$

This condition is typically satisfied in practice because:

- σ tends to be large in real-world domain shifts, as environments often differ significantly in their distributional properties.
- δ_{inv} is small because invariant features generalize well across environments and we can adopt proper realization of the predictor h to reduce δ_{inv}.
- $\epsilon(c)$ can be made small by ensuring Z^{inv} captures sufficient information from X.

This proves that under realistic conditions, predictors based on invariant representations achieve strictly lower worst-case out-of-distribution risk than those using raw representations. □

B Unimodal Networks

This section outlines the architecture of our unimodal networks and elaborates on the steps for generating unimodal representations, which serve as the foundation for subsequent causal analysis. To make a fair comparison, following established practices in recent work (Xiao et al., 2024; Mai et al., 2023b), we utilize pre-trained language models (He et al., 2021; Lan et al., 2020) to derive high-quality textual features. The following steps outline the language network's workflow for all downstream tasks

$$
\begin{aligned}
\hat{\boldsymbol{X}}_l &= \text{PLM}(\boldsymbol{U}_l;\ \theta_l) \in \mathbb{R}^{T_l\times d_l} \\
\boldsymbol{X}_l &= (\hat{\boldsymbol{X}}_l\boldsymbol{W}_{pro} + \boldsymbol{b}_{pro}) \in \mathbb{R}^{T_l\times d}
\end{aligned} \tag{10}
$$

where PLM indicates the pre-trained language model, $\boldsymbol{U}_l$ is the input token sequence and T_l represents the sequence length. $\boldsymbol{W}_{pro} \in \mathbb{R}^{d_l\times d}$ and $\boldsymbol{b}_{pro} \in \mathbb{R}^{1\times d}$ are trainable parameters that map the output dimensionality of the language network to the shared feature dimensionality d. In MSA, the acoustic and visual networks employ transformer encoders (Vaswani et al., 2017) and operate according to the following steps ($m \in \{a, v\}$):

$$
\begin{aligned}
\hat{\boldsymbol{X}}_m &= \text{Conv 1D}(\boldsymbol{U}_m;\ K_m) \in \mathbb{R}^{T_m\times d} \\
\boldsymbol{X}_m &= \text{Transformer}(\hat{\boldsymbol{X}}_m;\ \theta_m) \in \mathbb{R}^{T_m\times d}
\end{aligned} \tag{11}
$$

where $\boldsymbol{U}_m \in \mathbb{R}^{T_m\times d_m}$ is the extracted raw feature sequence (see Section E for the extraction derails), Conv 1D indicates the temporal convolution whose kernel size K_m is set to 3. Nota that for the CH-SIMS-v2 dataset, we use the same feature set as in previous works (Feng et al., 2024; Liu et al., 2022), which are feature vectors instead of sequences. Therefore, we simply use multi-layer perception networks as the unimodal networks for visual and acoustic modalities.

In the MHD and MSD tasks, to more effectively model humor-specific cues, we follow prior work (Hasan et al., 2021) by extracting an additional Humor-Centric Feature (HCF) from the language modality. This HCF serves as a fourth modality and is represented as $\boldsymbol{U_h} \in \mathbb{R}^{T_l\times d_h}$ (detailed in (Hasan et al., 2021)). In addition, each data sample in the MHD and MSD tasks includes both a target punchline segment and its preceding context. We merge the feature sequences of the punchline and context along the temporal axis to construct the unimodal input representations $\boldsymbol{U_m} \in \mathbb{R}^{T_m\times d_m}$ ($m \in \mathcal{M} = \{a, v, l, h\}$). The unimodal network for the HCF modality is analogous to the framework employed for the visual and acoustic modalities. The specific steps of the transformer-based unimodal networks for MHD and MSD are delineated as follows ($m \in \{a, v, h\}$):

$$
\begin{aligned}
\hat{\boldsymbol{X}}_m &= \text{Transformer}(\boldsymbol{U}_m;\ \theta_m) \in \mathbb{R}^{T_m\times d_m} \\
\boldsymbol{X}_m &= \text{Conv 1D}(\hat{\boldsymbol{X}}_m;\ K_m) \in \mathbb{R}^{T_m\times d}
\end{aligned} \tag{12}
$$

Finally, we employ a straightforward linear layer to fuse the language and HCF modalities, thereby reducing complexity for the following model stages:

$$
\boldsymbol{X}_l \longleftarrow \text{Linear}(\boldsymbol{X}_l \oplus \boldsymbol{X}_h;\ \theta_{lin}) \in \mathbb{R}^{T_l\times d} \tag{13}
$$

For all unimodal representations $\boldsymbol{X}_m \in \mathbb{R}^{T_m\times d}$, we perform mean pooling at the time dimension to obtain the final unimodal representations $X_m \in \mathbb{R}^{1\times d}$.

C Datasets

(1) **CMU-MOSI** (Zadeh et al., 2016): This dataset is a standard benchmark for MSA, encompassing more than 2,000 online video clips collected from the Internet. Each clip is labeled with a sentiment score on a -3 to 3 Likert scale, where 3 and -3 denote extreme positive and negative sentiments, respectively.

(2) **The OOD version of CMU-MOSI** (Sun et al., 2022): CMU-MOSI (OOD) is built using a modified simulated annealing algorithm (Aarts et al., 1987), which iteratively adjusts the test distribution. The resulting significant shifts in word–sentiment correlations relative to the training set establish it as a challenging benchmark for evaluating model robustness to distribution shifts in MSA.

(3) **CMU-MOSEI** (Zadeh et al., 2018b): The CMU-MOSEI dataset is a large-scale, widely-adopted benchmark for MSA, collected from online videos. Its key characteristics include: (I) Scale: over 22,000 video clips; (II) Source: more than 1,000 YouTube speakers and 250+ topics, randomly sampled; (III) Annotation: each clip has two labels—a six-class emotion category and a sentiment score ranging from -3 (strongly negative) to 3 (strongly positive). For the MSA task, our evaluation adopts the sentiment labels of CMU-MOSEI, which are consistent with the scale used in CMU-MOSI.

(4) **CH-SIMS-v2** (Liu et al., 2022): CH-SIMS-v2 serves as a Chinese MSA benchmark with the following characteristics: (I) Source: Videos collected from 11 scenarios (interviews, talk shows, films, etc.) to mimic real-world interaction; (II) Quality: Filtered to retain high-quality acoustic and visual streams; (III) Split: Partitioned into training, validation, and test sets in a 9:2:3 ratio, corresponding to 2,722, 647, and 1,034 segments respectively; (IV) Label Distribution: The training set contains 921 negative, 433 weakly negative, 232 neutral, 318 weakly positive, and 818 positive samples.

(5) **UR-FUNNY** (Hasan et al., 2019): Derived from TED talk videos involving 1,741 speakers, UR-FUNNY serves as a benchmark for multimodal humor detection (MHD). Each data sample includes a multimodal punchline segment and its preceding context segments, the latter being provided to support contextual modeling. The dataset is built by identifying punchlines via the laughter tag in transcripts. Video segments followed by laughter are treated as positive samples, while those without laughter form negative samples. It is partitioned into 7,614 training, 980 validation, and 994 test instances.

(6) **MUStARD** (Castro et al., 2019): The MUStARD dataset is designed for multimodal sarcasm detection (MSD), comprising video segments sourced from popular TV series including Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics. The collection contains 690 human-annotated segments, labeled as either sarcastic or non-sarcastic. Similar to UR-FUNNY, it provides contextual clues by incorporating both the target punchline and the preceding dialogue segments for each sample.

D Evaluation Metrics

For CMU-MOSI and CMU-MOSEI datasets, we adopt the following evaluation metrics: (1) **Acc7**: the accuracy of classifying sentiment scores into seven discrete classes; (2) **Acc2**: the binary accuracy for differentiating between positive and negative sentiments; (3) **F1 score**: a harmonic mean that balances precision and recall for binary sentiment classification; (4) **MAE**: the mean absolute error between model predictions and sentiment labels; and (5) **Corr**: the correlation coefficient reflecting the strength and direction of the relationship between predictions and sentiment labels. For Acc7, predictions are rounded to the nearest integer within the scale from -3 to 3. When calculating Acc2 and F1 score, neutral segments are not considered. And the neutral segments are included in the calculations of MAE, Corr, and Acc7. For the CH-SIMS-v2 dataset, we use Acc5, Acc3, Acc2, F1 score, MAE, and Corr as in previous works (Liu et al., 2022; Feng et al., 2024). For the MHD and MSD tasks, we report the binary accuracy (i.e., humorous or non-humorous, sarcastic or non-sarcastic) of the model.

E Feature Extraction Details

Regarding the **visual modality**, following previous approaches (Mai et al., 2023d; Xiao et al., 2024), Facet [1] is used to gather an array of visual attributes such as facial action units and facial landmarks for the MSA task. Facial feature extraction for the CH-SIMS-v2 dataset aligns with established practice (Feng et al., 2024; Liu et al., 2022), utilizing

[1] iMotions 2017. https://imotions.com/

Table 6: The unimodal feature dimensionality of different datasets.

	Language	Acoustic	Visual	HCF
CMU-MOSI	768	74	47	-
CMU-MOSI (OOD)	768	74	47	-
CMU-MOSEI	768	74	35	-
CH-SIMS-v2	768	25	177	-
UR-FUNNY	768	60	36	4
MUStARD	768	60	36	4

OpenFace (Baltrusaitis et al., 2018) to obtain measures such as 68 facial landmarks, 17 action units, head pose, head orientation, and eye gaze direction. For MHD and MSD, in line with prior approaches (Hasan et al., 2021; Mai et al., 2023a), OpenFace 2 (Baltrusaitis et al., 2018) is utilized for the extraction of facial action unit features as well as rigid and non-rigid facial shape parameters. For **acoustic modality**, COVAREP (Degottex et al., 2014) is used for the extraction of a sequence of acoustic features, including 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity, etc. For the CH-SIMS-v2 dataset, acoustic features are represented as 25-dimensional eGeMAPS low-level descriptors (LLD), extracted via OpenSmile (Eyben, 2010) at 16 kHz. For **language modality**, following state-of-the-art methods (Xiao et al., 2024), DeBERTa (He et al., 2021) and BERT (Devlin et al., 2019) are employed to learn informative language representations. For the CH-SIMS-v2 dataset, we follow prior works (Feng et al., 2024; Liu et al., 2022) and utilize BERT (Devlin et al., 2019) to obtain textual features. For MHD and MSD, following prior methods (Mai et al., 2023b,a), ALBERT (Lan et al., 2020) is adopted.

The input feature dimensionality for each modality is summarized in Table 6.

F Experimental Details

(1) **Hyperparameter Setting**: Our proposed CmIR is developed with PyTorch 1.13.1 on an NVIDIA RTX3090 GPU (CUDA 11.6). Training utilizes the AdamW optimizer (Loshchilov and Hutter, 2019). In line with prior work (Mai et al., 2023d), optimal hyperparameters are determined through an extensive random grid search of 50 iterations on the validation set. Using the identified best configuration, the model is retrained five times, and the final performance is averaged across these runs. The specific hyperparameter settings can be found in Table 7. Note that the modality-specific covariance matrix Σ_m in Eq. 3 is set as the identity matrix.

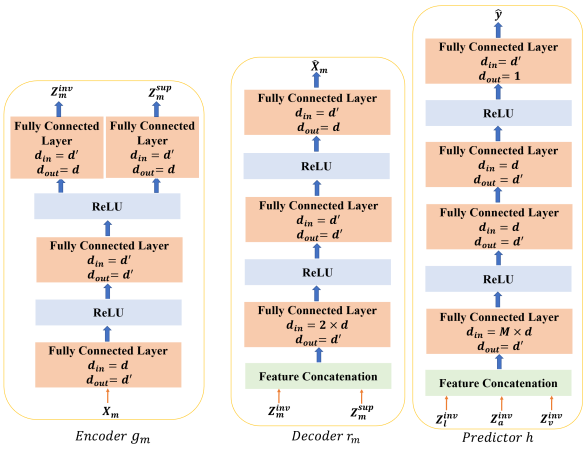


Figure 7: The structures of encoder, decoder, and predictor. d' represents the hidden dimensionality.

The structures of the predictive attention weight generator and the predictor are shown in Figure 7.

(2) **Protocol for the evaluation of noisy modalities**: To evaluate model robustness with noisy modalities, we produce corrupted features by employing the equation below:

$$X_m^n = (1 - NR) \cdot X_m + NR \cdot \mathcal{N} \qquad (14)$$

where X_m is the unimodal representation, NR is the noisy rate ranging from 0.1 to 0.7, $\mathcal{N}$ is the Gaussian noise data of mean **0** and variance **1**, and X_m^n is the noisy unimodal representation that is used to learn invariant representation. To simulate realistic noise, the described noise mixing process is applied to every modality across all samples. Perturbations are introduced at the feature level because input noise ultimately propagates to feature representations. This approach aligns with common practice in MAC, where a standardized feature set is typically used to ensure fairness and protect privacy, making feature-level noise injection both reasonable and practical. For a fair comparison, all baselines (Yuan et al., 2024; Mai et al., 2023c, 2024) are reproduced using the same training and testing protocols as our CmIR.

To further assess the robustness of CmIR under out-of-distribution (OOD) noise conditions (Table 4), we evaluate its performance against two additional corruption types: Laplace noise and random erasing noise (the latter simulates data loss by randomly zeroing out a subset of features). During this evaluation, each sample has an equal 50% chance of being corrupted either by Laplacian noise via Eq. 14 or by random feature dropout (missing), where the dropout rate is controlled by the noise ratio NR.

Table 7: Hyperparameter Settings of CmIR. MAE, MSE and BCE denote mean absolute error, mean square error and binary cross-entropy, respectively.

	CMU-MOSI	CMU-MOSEI	CH-SIMS-v2	MUStARD	UR-FUNNY	CMU-MOSI (OOD)
Loss Function	MSE	MSE	MAE	BCE	BCE	MSE
Batch Size	48	50	50	32	64	48
Learning Rate	1e-5	1e-5	5e-4	2e-5	7e-6	1e-5
Number of Environments K	1	5	2	4	2	4
Noise Intensity $\alpha^{(1)}$	0.1	1	0.1	0.1	0.1	0.1
Weight λ_1	0.1	0.1	0.01	0.05	0.001	0.01
Weight λ_2	0.001	0.1	0.01	0.01	0.01	0.01
Weight λ_3	0.05	0.01	0.1	0.05	0.01	0.01
Shared Dimensionality d	150	150	100	128	120	150
Hidden Dimensionality $d^{'}$	256	128	100	48	120	150
Weight α in Mutual Information Constraint	1	1	0.01	0.1	0.1	1

G Baselines

The causality-based baselines include:

(1) **Subject Causal Intervention (SuCI)** (Xu et al., 2025b): It introduces a simple yet effective causal intervention module designed to decouple the influence of subjects as unobserved confounders, thus obtaining unbiased predictions through true causal effects; (2) **CounterfactuaL mUltimodal sEntiment (CLUE)** (Sun et al., 2022): It leverages causal inference and counterfactual reasoning to prune away spurious direct textual influences, preserving only the genuine indirect multimodal effects, thereby strengthening generalization to out-of-distribution data; (3) **General dEbiAsing fRamework (GEAR)** (Sun et al., 2023): To enhance out-of-distribution robustness, it distinguishes robust features from biased ones, quantifies sample-level bias, and employs inverse probability weighting to de-emphasize highly biased samples; (4) **Multimodal Debiasing Framework (MulDeF)** (Huan et al., 2024): It integrates causal intervention with front-door adjustment and multimodal causal attention during the training phase. At inference, it applies counterfactual reasoning to mitigate both verbal and nonverbal biases, which enhances out-of-distribution generalization.; (5) **Attention-based Causality-Aware FusionAttention-based Causality-Aware Fusion (AtCAF)** (Huang et al., 2025): It learns causality-aware multimodal representations for sentiment analysis through a dedicated text debiasing module and counterfactual attention across modalities.

The baselines for handling noisy modality include:

(1) **Multimodal Boosting** (Mai et al., 2024): It is built upon multiple base learners organized in a boosting-like manner, with each learner addressing different facets of the multimodal data. It is further equipped with a contribution learning module that dynamically estimates the contribution and noise degree of individual learners; (2) **Complete Multimodal Information Bottleneck (C-MIB)** (Mai et al., 2023c): It adopts the information bottleneck principle to eliminate redundancy and noise from both unimodal and multimodal features, thus establishing it as a baseline for handling noisy modalities.; (3) **Two-stage Modality Denoising and Complementation (TMDC)** (Zhuang et al., 2025): Its training process involves two distinct stages. The first, the intra-modality denoising stage, aims to enhance representational robustness by using denoising modules to obtain clean modality-specific and shared representations from complete data, thereby reducing noise interference. The second, the inter-modality complementation Stage, utilizes these representations to address modality absence through cross-modal compensation.

The additional baselines for MSA include:

(1) **Information-Theoretic Hierarchical Perception (ITHP)** (Xiao et al., 2024): Its design is grounded in the information bottleneck principle, where one modality is designated as the core, while others act as detectors within the information pathway to distill and refine the information flow; (2) **Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM)** (Yu et al., 2021): This approach employs a self-supervised strategy to infer sentiment labels for individual modalities by leveraging the global labels of multimodal samples, thereby learning more discriminative unimodal representations; (3) **Disentangled-Language-Focused Model (DLF)** (Wang et al., 2025): It introduces a feature disentanglement module to separate shared and specific information across modalities. The process is further refined by four geometric measures to reduce redundancy and prioritize language-focused features. A language-targeted attractor is also designed to enhance lan-

guage representations using complementary information from other modalities; (4) **Gradient and Structure Consistency** (**GSCon**) (Shi et al., 2025): It proposes a balanced gradient direction that aligns each modality's optimization direction to ensure unbiased convergence, and aligns the spatial structure of samples in different modalities to avoid the interaction noise caused by multimodal alignment; (5) **Diffusion Bridge** (Lee et al., 2025): It directly reduces the modality gap by leveraging denoising diffusion probabilistic models; (6) **Enhanced Dynamic Emotion Experts** (**EMOE**) (Fang et al., 2025): The framework comprises a mixture of modality experts that dynamically adjust modality importance based on input features, coupled with a unimodal distillation mechanism to preserve the predictive capacity of individual modalities within the fused representation; (7) **Contrastive FEature DEcomposition** (**ConFEDE**) (Yang et al., 2023): It enhances multimodal representations by jointly performing contrastive learning and contrastive decomposition of features; (8) **Kolmogorov–Arnold Network with Multimodal Clean Pareto KAN-MCP** (Luo et al., 2025a): It combines interpretable cross-modal modeling via KANs with feature denoising and compression based on DRD-MIB, yielding discriminative multimodal inputs while alleviating modality imbalance; (9) **Multimodal Adaptation Gate BERT/ALBERT** (**MAG-BERT/MAG-ALBERT**) (Rahman et al., 2020): It incorporates visual and acoustic information into BERT/ALBERT through a dedicated multimodal adaptation gate; (10) **Acoustic Visual Mix-up Consistent** (**AV-MC**) (Liu et al., 2022): This method utilizes modality mix-up to augment visual and acoustic representations, thereby strengthening their contribution to sentiment analysis; (11) **Knowledge-Guided Dynamic Modality Attention Fusion** (**KUDA**) (Feng et al., 2024): It guides the multimodal fusion process with external emotional knowledge, dynamically selecting the dominant modality and adjusting the weighting of all modalities; (12) **MISA** (Hazarika et al., 2020): It decomposes each unimodal input into modality-invariant and modality-specific components, which are subsequently fused for final prediction; (13) **MultiModal InfoMax** (**MMIM**) (Han et al., 2021): In MMIM, representation learning is enhanced by maximizing the mutual information both between unimodal features and between multimodal representations at different levels and their unimodal counterparts; (14) **Feature-Disentangled Multimodal Emotion Recognition** (**FDMER**) (Yang et al., 2022): It learns the common and private feature representations for each modality, which achieves the modality consistency and disparity constraints by designing tailored losses for modality-invariant and modality-specific subspaces.

The additional baselines for MHD and MSD include:

(1) **Multimodal Global Contrastive Learning** (**MGCL**) (Mai et al., 2023b): MGCL applies supervised contrastive learning to multimodal representations, employing diverse augmentation strategies to construct positive and negative sample pairs for each representation; (2) **Multimodal Correlation Learning** (**MCL**) (Mai et al., 2023a): MCL formulates a supervised correlation learning objective that preserves modality-specific characteristics while fostering a more discriminative joint embedding space; (3) **Multimodal Multitask Interaction Learning** (**MIL**) (Zhang et al., 2024): It performs joint sarcasm and sentiment detection, integrating a cross-modal target attention mechanism to align textual with visual/acoustic content and a multimodal interaction module to model the shared and distinct patterns of both tasks; (4) **Decoupled Multimodal Distillation** (**DMD**) (Li et al., 2023): DMD enables adaptive cross-modal knowledge transfer by decoupling unimodal features into modality-irrelevant and modality-exclusive subspaces, followed by a specialized graph distillation unit to handle each subspace effectively; (5) **Humor Knowledge Enriched Transformer** (**HKT**) (Hasan et al., 2021): HKT incorporates humor-centric features as external knowledge to resolve the ambiguity and leverage the subtle sentiment cues within the language modality; (6) **Multimodal Transformer** (**MulT**) (Tsai et al., 2019a): MulT leverages stacked cross-modal transformers to project and align source modalities to a target modality, effectively mitigating the inter-modal gap.

H Comparison between Different Invariance Constraints

In Section 4.2, we propose two different methods to realize invariance constraint. The first one is a standard way to compute and minimize the conditional distribution distance (we directly minimize unimodal predictions from different environments for regression tasks, and use KL-divergence to min-

Table 8: Further discussions on different invariance constraints.

	CMU-MOSI					CMU-MOSEI					MUStARD
	Acc7	Acc2	F1 score	MAE	Corr	Acc7	Acc2	F1 score	MAE	Corr	Acc
Standard	48.4	**90.1**	**90.1**	**0.616**	**0.855**	**55.2**	87.4	87.3	0.523	0.781	79.7
Ours	**49.8**	89.6	89.5	**0.616**	0.853	55.1	**87.8**	**87.7**	**0.513**	**0.793**	**80.0**

imize conditional distribution distance for classification tasks), which needs to implement a unimodal predictor for each modality and generate unimodal prediction. The second one is to directly minimize the distance between different causal invariant representations extracted under different environments, which is adopted in our CmIR. Here we provide a comparison between these two strategies, and the results are shown in Table 8. From Table 8 we can infer that these two strategies actually yield comparable outcomes, likely because although the first method can more directly evaluate the differences between conditional distributions, it introduces a unimodal predictor. However, when the unimodal predictor is insufficiently trained, it may introduce evaluation errors. Therefore, we adopt the second approach, which enables a more rigorous constraint without introducing a unimodal predictor, thereby reducing model complexity.

I Model Complexity Analysis

The proposed CmIR does not design complex multimodal fusion to explore sufficient inter-modal interactions, but merely introduces an encoder and a decoder to learn invariant modality representations for each modality. Each encoder/decoder only has a few layers. Therefore, the model complexity of the proposed CmIR is acceptable. To verify our claim, we compare the model complexity of CmIR with competitive baselines on the MUStARD (Castro et al., 2019) and CMU-MOSI (Zadeh et al., 2016) datasets. As shown in Table 9, on MUStARD, **CmIR outperforms competitive baselines with minimal number of parameters**, demonstrating the effectiveness of the proposed framework and indicating that the performance improvement of CmIR dose not results from the increase in the number of parameters. On CMU-MOSI, we evaluate the FLOPs and number of parameters of our CmIR. As shown in Table 9, CmIR has significantly fewer parameters than EMOE and GSCon, slightly more than Diffusion Bridge, while its FLOPs are lower than all baselines, indicating the high efficiency of our method.

J Significant Test

In this section, we provide the results of significant test between the proposed CmIR and a strong baseline ITHP (Xiao et al., 2024) using t-test on the CMU-MOSI and CMU-MOSEI datasets. We run each model for ten times under different random seeds to gather the results. As shown in Table 10, the t-test results indicate that CmIR has a statistically significant difference with ITHP (Xiao et al., 2024) for all evaluation metrics except the Corr metric ($p < 0.05$ denotes a statistically significant difference), suggesting that the improvement of CmIR is significant.

K Probing Experiments for Z^{inv} and Z^{spu}

Orthogonality is only a practical proxy for mutual information minimization. To strengthen the evidence that Z^{inv} and Z^{spu} are well separated and contain different information, we have added two probing experiments: (a) training a predictor to predict the environment label from Z^{inv} or Z^{spu} alone. Since the difference between different environments lies in the noise level (noise coefficient $\alpha^{(e)}$), we directly train an additional fusion network and predictor that use Z^{inv} / Z^{spu} to predict the noise coefficient $\alpha^{(e)}$. Finally, we use MSE to evaluate the effectiveness of the prediction (lower accuracy from Z^{inv} indicates better invariance); (b) Predicting the sentiment label from each component separately. As shown in Table 11, Z^{inv} yields high MSE in environment prediction but high accuracy in sentiment prediction, while Z^{spu} does the opposite, confirming semantic separation. Notably, Z^{spu} only relies on label bias for prediction (predicting all samples as positive yields an accuracy of 62.8% which is exactly the accuracy of the prediction from Z^{spu}), suggesting it contains non-causal information.

L Sensitivity Analysis of Noise Intensity

In this section, we have conducted new experiments on CMU-MOSI (OOD) and provided the sensitivity analysis of noise intensity $\alpha^{(1)}$ in Table 12. As shown in Table 12, when the value of $\alpha^{(1)}$ is small,

Table 9: Model complexity analysis on the MUStARD and CMU-MOSI datasets.

MUStARD			CMU-MOSI			
Model	Acc2	Parameters (M)	Model	Acc2	Parameters (M)	FLOPs (G)
HKT (Hasan et al., 2021)	76.47	17.10M	EMOE (Fang et al., 2025)	84.8	317.30M	10.02
MCL (Mai et al., 2023a)	77.94	13.83M	GSCon (Shi et al., 2025)	88.1	248.52M	12.22
MGCL (Mai et al., 2023b)	77.94	14.28M	Diffusion Bridge (Lee et al., 2025)	86.9	**185.46M**	11.52
CmIR	**80.00**	**13.44**	CmIR	**89.6**	187.51M	8.95

Table 10: **Paired t-test analysis between CmIR and ITHP (Xiao et al., 2024).**

	CMU-MOSI					CMU-MOSEI				
	Acc7	Acc2	F1	MAE	Corr	Acc7	Acc2	F1	MAE	Corr
ITHP	0.002	0.004	0.004	4.98e-4	0.439	3.65e-5	5.73e-4	7.78e-4	1.12e-4	0.566

CmIR performs poorly. This is because when the imposed noise is small, its impact on modality representations is limited, making it difficult to simulate diverse environments for learning effective causal invariant representations. Moreover, when $\alpha^{(1)}$ fluctuates within a large range (from 1e-7 to 1), the model maintains good performance, demonstrating the robustness of CmIR. Additionally, when $\alpha^{(1)} = 1$, the model achieves stronger performance than the default setting ($\alpha^{(1)} = 0.1$), indicating the performance potential of CmIR (better results can be obtained through more careful hyperparameter tuning). Compared with the number of environments K (see Figure 5 (d)), $\alpha^{(1)}$ has a greater impact and should be set to a relatively large value.

Table 11: **Probing results on CMU-MOSEI.**

Component	Environment Prediction	Sentiment prediction				
	MSE	Acc7	Acc2	F1	MAE	Corr
Z^{inv}	2.04	55.1	87.8	87.7	0.513	0.793
Z^{spu}	0.87	41.4	62.8	48.5	0.841	0.001

Table 12: **Sensitivity Analysis of Noise Intensity $\alpha^{(1)}$ on CMU-MOSI (OOD).**

$\alpha^{(1)}$	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
Acc7	43.5	45.8	47.5	47.0	47.3	48.0	48.8	48.5	49.0
Acc2	79.7	82.8	84.1	83.6	84.0	84.4	83.9	84.4	84.6
F1 score	79.7	82.7	83.9	83.5	84.1	84.3	83.9	84.4	84.6