Plug-and-play Feature Causality Decomposition for Multimodal Representation Learning

Ye Liu **Zihan Ji** **Hongmin Cai***
School of Future Technology
South China University of Technology
Guangzhou, China 511442
{yliu03, hmcai}@scut.edu.cn ftjizihan@mail.scut.edu.cn

Abstract

Multimodal representation learning is critical for a wide range of applications, such as multimodal sentiment analysis. Current multimodal representation learning methods mainly focus on the multimodal alignment or fusion strategies, such that the complementary and consistent information among heterogeneous modalities can be fully explored. However, they mistakenly treat the uncertainty noise within each modality as the complementary information, failing to simultaneously leverage both consistent and complementary information while eliminating the aleatoric uncertainty within each modality. To address this issue, we propose a plug-and-play feature causality decomposition method for multimodal representation learning from causality perspective, which can be integrated into existing models with no affects on the original model structures. Specifically, to deal with the heterogeneity and consistency, according to whether it can be aligned with other modalities, the unimodal feature is first disentangled into two parts: modality-invariant (the synergistic information shared by all heterogeneous modalities) and modality-specific part. To deal with complementarity and uncertainty, the modality-specific part is further decomposed into unique and redundant features, where the redundant feature is removed and the unique feature is reserved based on the backdoor-adjustment. The effectiveness of noise removal is supported by causality theory. Finally, the task-related information, including both synergistic and unique components, is further fed to the original fusion module to obtain the final multimodal representations. Extensive experiments show the effectiveness of our proposed strategies.

1 Introduction

Multimodal representation learning is the basis of the downstream tasks, such as multimodal sentiment analysis [39; 13; 9; 15] and text-image classification [18; 11]. Due to the feature heterogeneity among different modalities [36], the consistent and complementary information among modalities [12] are crucial for effective multimodal representation learning. The consistent information refers to common features that present in multiple modalities, facilitating alignment of modalities within a common representation space. On the other hand, the complementary information refers to distinctive features presenting in only one modality but is still useful for the overall understanding. It enriches multimodal models, enabling them to learn more comprehensive and nuanced representations. Numerous existing methods have been proposed to effectively explore them. For example, [13] captures the consistent information via maximizing mutual information among different modalities, thereby improving multimodal fusion performances. [20] applies the contrastive learning to enhance

*The corresponding author.

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

the feature representation capability of consistent information on both label-level and instance-level. [5] performs dynamic late fusion by predicting the generalization error upper bound of each modality, exploiting the heterogeneous complementary information in each modality.

However, due to the variations in data acquisition processes and sensor characteristics, multimodal data contain distinct types of uncertainty noise in each modality. We carry out a qualitative analysis with the help of structural causal model (SCM) [25] to explain it. Figure 1 describes the SCM of multimodal data of a single modality. The consistent and complementary information are gathered from the task-related factor, from where the synergistic and unique features are to be constructed respectively. The synergistic feature is the shared consistent information among different modalities, which is modality-invariant. And the unique feature is the complementary information that is modality-specific. However, due to the modality uncertainty noise factor resulted from the aforementioned cues, the redundant task-agnostic information is also mixed in the modality-specific feature, which would affect the construction of unique feature. For example, in image-text task, the task-related parts are the ones that contribute to the final task prediction. The synergistic parts are both included in the image and the text modalities, which jointly contribute to the final task prediction. The unique parts are the semantic contents that only appear in image modality (or text modality) and don't appear in text modality (or image modality). Both synergistic and unique factors are task-related and derived from the task-related factor. The redundant part is the uncertainty noise within one modality, such as the typos in text modality or quality of image, which is derived from the noise factor. The multimodal data store these information in their different format (the specific factor), but they are still associated in high-level semantic (the synergistic factor). Existing methods often overlook this unimodal uncertainty noise, mistakenly treating it as complementary information, which hampers multimodal fusion performance [11]. Moreover, several researches have shown that removing uncertainty in unimodal features helps to obtain more accurate and robust multimodal representations. For example, [11] assumes the unimodal features are sampled from a multivariate Gaussian distribution and tries to remove the aleatoric uncertainty by estimating their latent distributions. [38] applies causal intervention on the amplitude information after the Fourier transformation of different modality samples, and enforces model on high-level semantic information. Nevertheless, these methods typically focus solely on consistent, complementary, or uncertain information, or a combination of two of these aspects, with very few approaches effectively capturing all three types simultaneously.

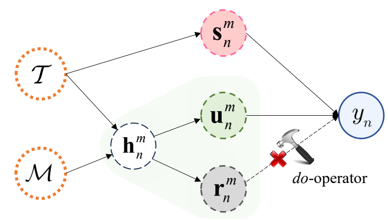
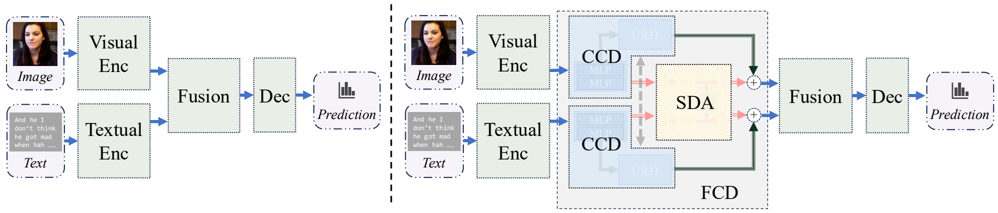


Figure 1: Structural causal model for the n-th multimodal data of the m-th modality. $\mathcal{T}$: task-related factor. $\mathcal{M}$: modality uncertainty noise factor. $\mathbf{s}_n^m$: synergistic feature. $\mathbf{h}_n^m$: modality-specific information. $\mathbf{u}_n^m$: unique feature. $\mathbf{r}_n^m$: modality-specific task-agnostic information. y_n: task annotation.

Inspired by [24; 45], we propose a plug-and-play Feature Causality Decomposition (FCD) method to address the heterogeneity and aleatoric uncertainty of multimodal data. Ideally, the task-related information within each modality should be fully utilized to perform the multimodal representation learning, whilst the task-agnostic information should be eliminated. Thus, FCD is designed to take the original unimodal features as inputs and decompose them into their synergistic, unique and redundant components. To achieve this, FCD first employs Causality Components Decomposition (CCD) module on each unimodal feature to decompose it into the modality-invariant and modality-specific parts. The modality-invariant part plays a role of synergistic component, which is capable of capturing the consistent information shared across heterogeneous modalities. FCD employs another module named Synergistic Distribution Alignment (SDA) to narrow the space difference and align the synergistic component from all modalities via a parameter-sharing MLP constrained by Sinkhorn divergence [10]. Then, CCD further extracts the unique component from the modality-specific part based on the backdoor-adjustment [25]. We theoretically [16; 25] prove that the unimodal uncertainty noise mixed in the modality-specific part can be removed via the backdoor-adjustment under some specific conditions. Moreover, contrastive loss among the redundant components from different modalities is utilized to constrain the modality-specific characteristic, effectively handling the uncertainties that vary across modalities. Different from existing multimodal learning methods [9; 13; 44; 45], FCD is plug-and-play that is more generalized and achieves the theoretical support from causality and probability perspectives. Extensive experiments on 9 existing multimodal methods and 5 multimodal datasets prove the effectiveness of FCD.



(a) Intermediate Fusion Pipeline (b) Intermediate Fusion Pipeline with FCD

Figure 2: Take image and text modalities as examples. “Enc” and “Dec” stand for encoder and prediction head, respectively. Arrows are the forward paths. (a) The original pipeline of multimodal intermediate fusion model. (b) The multimodal intermediate fusion pipeline involving FCD module, which takes unimodal features as inputs and outputs the synergistic (pink double-line shaft arrows), unique (green triple-line shaft arrows) and redundant (gray dashed arrows) components.

2 Related Works

2.1 Multimodal Representation Learning

Multimodal representation learning has gained increasing attention with the rapid advancement of social media and multimedia technologies [31; 11]. Moreover, feature extraction plays a crucial role in multimodal representation learning, and numerous studies have been conducted. Recent efforts have explored various strategies for effective multimodal learning. [11] mitigate aleatoric uncertainty within each modality via latent distribution modeling, followed by dynamic integration and VIB-based fusion [1]. [20] performs token-level fusion with contrastive learning to align cross-modal representations. [13] reduce modality discrepancy via mutual information minimization and introduce CPC [28] to preserve modality-invariant features. [9] treat fusion as a progressive process and employ a hierarchical contrastive framework to retain semantic consistency across fusion stages. These works hardly considered the heterogeneity and aleatoric uncertainty of the multimodal data simultaneously. In addition to these works, many works on other tasks, such as multimodal learning with miss-modality [35], semantic segmentation [44] and multimodal domain generalization [8], focused on the modality-specific/-invariant features but ignored the unimodal uncertainty noise. Recently, [45] noticed the uncertainty noise within each modality and proposed to remove it through cross-attention mechanism. Unlike [45], we provide a simpler yet more generalized plug-and-play module whose effectiveness is also supported by causality and probability theory.

2.2 Causal Inference in Deep Learning

Causal inference [25] has attracted massive attention in multimodal tasks, where causality is defined as the relation between an event and the cause that gives rise to it [38]. It improves the robustness and generalization of models by debiasing the specific spurious correlation [6; 15; 23]. [23] used the backdoor and frontdoor causal intervention to eliminate the noise and the spurious correlation between input multimodal data and the ground truth in visual question answering task. [6] proposed to remove the influence of word frequency in textual modality by backdoor-adjustment. Besides, they also argued that the authenticity of the news should not be inferred only based on image features, and proposed to utilize the indirect effect of the image feature via counterfactual inference. [38] considered the modality-specific characteristic as the amplitude information of the Fourier transformation, and proposed to remove the modality noise by perturbing the amplitude component and generate a counterfactual sample to perform counterfactual inference.

3 Preliminaries

Given a multimodal dataset $\mathcal{D}$ with N samples in M modalities, and each sample pair is denoted as $(\mathbf{X}_n^m, y_n)$, where $\mathbf{X}_n^m$ is the n-th data sample in the m-th modality, y_n is the label of the n-th sample. $y_n \in \mathbb{R}$ is a real-valued number in regression tasks, while it is a discrete number, *i.e.*, $y_n \in \{1, \cdots, K\}$ with K being the number of clusters in classification tasks. In this paper, we specifically focus on multimodal intermediate fusion model, an example with visual and textual modalities is shown in Figure 2 (a). Multimodal intermediate fusion usually involves three steps, which can be formulated as follows.

$$\hat{y}_n = \mathcal{G}(\mathbf{z}_n; \theta_{\text{Dec}}), \quad \mathbf{z}_n = \mathcal{F}_{\text{Fuse}}(\mathbf{z}_n^1, \cdots, \mathbf{z}_n^M; \theta_{\text{Fuse}}) \in \mathbb{R}^d$$
$$\text{s.t.} \quad \mathbf{z}_n^m = \mathcal{E}^m(\mathbf{X}_n^m; \theta_{\text{Enc}}^m) \in \mathbb{R}^{d^m}, m = \{1, \cdots, M\} \tag{1}$$

where $\mathcal{E}^m(\cdot; \theta_{\text{Enc}}^m)$ is the encoder for the m-th modality with parameter θ_{Enc}^m, $\mathcal{F}_{\text{Fuse}}(\cdot; \theta_{\text{Fuse}})$ is the fusion function with parameters θ_{Fuse}, and $\mathcal{G}(\cdot; \theta_{\text{Dec}})$ is the predictive function with parameters θ_{Dec}. The optimization function of the regression and classification tasks is usually represented as:

$$\text{Regression tasks: } \mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{n=1}^{N} |y_n - \hat{y}_n| + \mathcal{L}_{\text{reg}}$$
$$\text{Classification tasks: } \mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{reg}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{I}_{y_n=k} \log \hat{y}_{nk} + \mathcal{L}_{\text{reg}} \tag{2}$$

where the indicator function $\mathbb{I}_{y_n=k} = 1$ if and only if y_n equals to k, otherwise $\mathbb{I}_{y_n=k} = 0$. $\hat{y}_n$ is the prediction of the n-th sample. $\hat{y}_{nk}$ is the predicted probability that the n-th sample belongs to the k-th category, and $\mathcal{L}_{\text{reg}}$ is the regularization term. Different $\mathcal{L}_{\text{reg}}$ is employed in different existing works. For example, a series of contrastive learning loss terms at different fusion phases are proposed in [9].

4 Methodology

The overview of our proposed method, Feature Causality Decomposition (FCD), is shown in Figure 2 (b). FCD is a plug-and-play module, which can be integrated into any intermediate fusion model for multimodal learning defined in Section 3. It is located between the unimodal encoder and the original fusion module, which takes the unimodal features $\{\mathbf{z}_n^m\}_{m=1}^M$ as input and feeds the processed features $\{\tilde{\mathbf{z}}_n^m\}_{m=1}^M$ to the original fusion module.

FCD contains two main submodules for each modality, *i.e.*, Causality Components Decomposition (CCD) and Synergistic Distribution Alignment (SDA). CCD first separates the unimodal feature $\mathbf{z}_n^m$ into three aforementioned components, one of which can be aligned with the same part from other modalities (*i.e.*, the synergistic component, pink double-line shaft arrows in Figure 2 (b), denoted as $\mathbf{s}_n^m$). The modality-specific task-related unique component is illustrated as green triple-line shaft arrows in Figure 2 (b) and denoted as $\mathbf{u}_n^m$. The last one is the redundant component (the gray dashed arrows in Figure 2 (b), denoted as $\mathbf{r}_n^m$), which is the modality uncertainty noise. Then, $\mathbf{s}_n^m$ is aligned by the SDA module. Finally, FCD aggregates the two task-related component, *i.e.*, synergistic and unique features, and feeds the aggregated feature $\tilde{\mathbf{z}}_n^m$ to the original fusion module $\mathcal{F}_{\text{Fuse}}$.

4.1 Causality Components Decomposition Module

The illustration of CCD is shown in Figure 3. Inspired by [45], CCD first disentangles $\mathbf{z}_n^m$ in Equation (1) to the modality-specific part $\mathbf{h}_n^m$ and the modality-invariant part $\mathbf{s}_n^m$ via two MLPs $\mathcal{F}_{\text{h}}^m$ and $\mathcal{F}_{\text{s}}^m$:

$$\mathbf{h}_n^m = \mathcal{F}_{\text{h}}^m(\mathbf{z}_n^m; \theta_{\text{h}}^m), \quad \mathbf{s}_n^m = \mathcal{F}_{\text{s}}^m(\mathbf{z}_n^m; \theta_{\text{s}}^m) \tag{3}$$

where θ_{h}^m and θ_{s}^m are parameters of the $\mathcal{F}_{\text{h}}^m : \mathbb{R}^{d^m} \to \mathbb{R}^{d^m}$ and $\mathcal{F}_{\text{s}}^m : \mathbb{R}^{d^m} \to \mathbb{R}^{d^m}$, respectively.

As discussed in Section 1, current methods may mistakenly treat features with uncertainty noise as the part of modality-specific information $\mathbf{h}_n^m$. In other words, $\mathbf{h}_n^m$ contains both modality-specific task-related component $\mathbf{u}_n^m$ (complementary information) and redundant component $\mathbf{r}_n^m$. Therefore, the core problem is how to reserve $\mathbf{u}_n^m$ and eliminate $\mathbf{r}_n^m$ from the $\mathbf{h}_n^m$ simultaneously. From the causality perspective, $\mathbf{u}_n^m$ is confounded by $\mathbf{h}_n^m$. As shown in Figure 1, there exists a backdoor path between $\mathbf{u}_n^m$ and y_n, which is confounded by the modality-specific feature $\mathbf{h}_n^m$: $\mathbf{u}_n^m \leftarrow \mathbf{h}_n^m \rightarrow \mathbf{r}_n^m \rightarrow y_n$. To eliminate the causal effect of the redundant component $\mathbf{r}_n^m$ in the backdoor

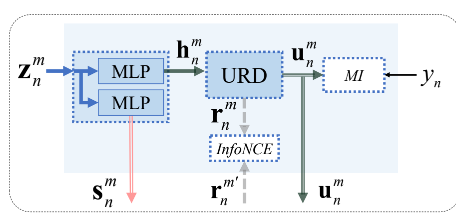


Figure 3: The illustration of the proposed Causality Components Decomposition (CCD) module.

path, backdoor-adjustment is commonly used [25; 23; 15], which is usually formed as the $do(\cdot)$. It performs causal intervention on the confounder factor to remove its spurious causal effect. In CCD, backdoor-adjustment is also employed to remove the causal effect of the confounder in backdoor path.

Theorem 4.1. *Given $\mathbf{h}_n^m$ extracted by $\mathcal{F}_h^m$, the conditional mutual information (CMI) between $\mathbf{u}_n^m$ and y_n is defined as $I(\mathbf{u}_n^m; y_n|\mathbf{h}_n^m)$. Assuming $\mathbf{u}_n^m$ is extracted from $\mathbf{h}_n^m$ by a measure-preserving bijective function $\mathcal{F}_{mb}^m$, then maximizing the expectation of conditional mutual information for all possible $\mathbf{h}_n^m$, i.e., $\mathbb{E}_{P(\mathbf{h}_n^m)}[I(\mathbf{u}_n^m; y_n|\mathbf{h}_n^m)]$, is equivalent to maximizing the mutual information $I(do(\mathbf{u}_n^m); y_n)$, where do-operator $do(\cdot)$ is the employed backdoor-adjustment, $\mathbb{E}$ is the expectation, and P is the probability distribution function.*

The proof [16; 25; 21] of Theorem 4.1 can be found in Appendix Section A. Theorem 4.1 indicates that if the extraction function $\mathcal{F}_{\text{mb}}^m$ is measure-preserving bijective, then effectively extracting the task-related feature $\mathbf{u}_n^m$ from the modality-specific feature $\mathbf{h}_n^m$ while separating out the modality noise $\mathbf{r}_n^m$ via backdoor adjustment is equivalent to maximizing the mutual information between the annotation and $\mathbf{u}_n^m$ after causal intervention. When $\mathbf{u}_n^m$ is extracted from $\mathbf{h}_n^m$ by $\mathcal{F}_{\text{mb}}^m$, *i.e.*, $\mathbf{u}_n^m = \mathcal{F}_{\text{mb}}^m(\mathbf{h}_n^m)$, CCD is able to extract $\mathbf{u}_n^m$ from $\mathbf{h}_n^m$ by optimizing the following:

$$\begin{aligned}
&\min \mathcal{L}_{\text{CMI}} \overset{\text{def.}}{=} \frac{-1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}_{P(\mathbf{h}_n^m)}[I(\mathbf{u}_n^m; y_n|\mathbf{h}_n^m)] \\
&\text{s.t. } \; I(\mathbf{u}_n^m; y_n|\mathbf{h}_n^m) = \mathbb{E}_{P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m)}\left[\log \frac{P(\mathbf{u}_n^m, y_n|\mathbf{h}_n^m)}{P(\mathbf{u}_n^m|\mathbf{h}_n^m)P(y_n|\mathbf{h}_n^m)}\right]
\end{aligned} \tag{4}$$

However, the true distribution of $\mathbf{h}_n^m$ is not assessable, which impedes the calculation of Equation (4). More importantly, directly optimizing Equation (4) without adjustment ignores the existing backdoor path, which would spuriously bring task-agnostic information towards annotation prediction. Based on Theorem 4.1, we can transform Equation (4) to Equation (5) when $\mathcal{F}_{\text{mb}}^m$ is a measure-preserving bijective function and backdoor-adjustment is applied to ensure the mixed uncertainty noise can be removed.

$$\begin{aligned}
\min \mathcal{L}_{\text{MI}} &\overset{\text{def.}}{=} \frac{-1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} I(do(\mathbf{u}_n^m); y_n) \\
&= \frac{-1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}_{P(do(\mathbf{u}_n^m), y_n)}\left[\log \frac{P(do(\mathbf{u}_n^m), y_n)}{P(do(\mathbf{u}_n^m))P(y_n)}\right]
\end{aligned} \tag{5}$$

According to [28; 22; 9], CCD applies InfoNCE loss as its lower bound in practice. Hence, we can transform Equation (5) to Equation (6) to perform the backdoor-adjustment and reserve deconfounded modality-specific task-related feature $do(\mathbf{u}_n^m)$.

$$\begin{aligned}
\mathcal{L}_{\text{MI}} \approx \mathcal{L}_{\text{InfoNCE}} &= \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathcal{L}_{\text{InfoNCE}}(\tilde{\mathbf{u}}_n^m, \tilde{\mathbf{y}}_n) \\
&= \frac{-1}{2NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[\log \frac{e^{\cos(\tilde{\mathbf{u}}_n^m, \tilde{\mathbf{y}}_n)/\tau}}{\sum_{n'=1}^{N_c} e^{\cos(\tilde{\mathbf{u}}_n^m, \tilde{\mathbf{y}}_{n'})/\tau}} + \log \frac{e^{\cos(\tilde{\mathbf{u}}_n^m, \tilde{\mathbf{y}}_n)/\tau}}{\sum_{n'=1}^{N_c} e^{\cos(\tilde{\mathbf{u}}_{n'}^m, \tilde{\mathbf{y}}_n)/\tau}}\right]
\end{aligned} \tag{6}$$

where $\tilde{\mathbf{u}}_n^m \in \mathbb{R}^d$ and $\tilde{\mathbf{y}}_n \in \mathbb{R}^d$ are linearly projected from $do(\mathbf{u}_n^m)$ and y_n with normalization, respectively. τ is the temperature parameter. And $\cos(\cdot, \cdot)$ is the cosine between two vectors.

On the basis of Theorem 4.1, effective extraction of $\mathbf{u}_n^m$ from $\mathbf{h}_n^m$ whilst separating $\mathbf{r}_n^m$ apart requires that the extraction function $\mathcal{F}_{\text{mb}}^m$ is a measure-preserving bijective function. According to [46; 45], we propose a *binary disentangle* module Unique Redundant Decompose (URD), which is guaranteed to satisfy the measure-preserving bijective condition.

$$\begin{aligned}
&\mathbf{u}_n^m = \mathcal{F}_{\text{URD-f}}^m\left(\sum_{k=1}^{3} w_k^m \mathbf{u}_{n(k)}^m; \theta_{\text{URD-f}}^m\right), \quad \mathbf{r}_n^m = \mathbf{h}_n^m - \mathbf{u}_n^m \\
&\text{s.t. } \; \mathbf{u}_{n(k)}^m = \mathcal{F}_{\text{URD-d}}^m(\mathbf{h}_{n(k)}^m; \theta_{\text{URD-d}}^m)
\end{aligned} \tag{7}$$

where the scaler $w_k^m \in \mathbb{R}$ is linearly projected from the dimensional concatenation of $\mathbf{r}_{n(k)}^m$ and $\mathbf{h}_{n(k)}^m$. $\mathcal{F}_{\text{URD-d}}^m : \mathbb{R}^{d^m} \to \mathbb{R}^{d^m}$ is a MLP for decomposition with parameter $\theta_{\text{URD-d}}^m$. $\mathcal{F}_{\text{URD-f}}^m : \mathbb{R}^{d^m} \to \mathbb{R}^{d^m}$ is a feed foward net with parameter $\theta^m_{\text{URD-f}}$ and $\mathbf{h}^m_{n(k)}$ is recursively deduced as:

$$\mathbf{h}^m_{n(k)} = \begin{cases} \mathbf{h}^m_n, k = 1 \\ \mathbf{h}^m_n - \sum_{i=1}^{k-1} w^m_i \mathbf{u}^m_{n(i)}, k > 1 \end{cases} \tag{8}$$

Practically, we apply SVD on the weight matrices of $\mathcal{F}^m_{\text{URD-f}}$ and $\mathcal{F}^m_{\text{URD-d}}$, where the 0 diagonal elements of the singular value matrix Σ are set to 10^{-5}. The pseudo code is in Appendix Section B.

On the other hand, the modality-specific task-agnostic component $\mathbf{r}^m_n$ should have significant difference with $\mathbf{r}^{m'}_n$ $(m' \neq m)$ to constrain the modality discrimination of $\mathbf{h}^m_n$. To achieve it, we employ InfoNCE [28] to perform contrastive learning between the redundant features from different modalities of the same sample as an unsupervised optimization target.

$$\min \mathcal{L}_{\text{Dis}} \overset{\text{def.}}{=} \frac{-1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \log \frac{e^{\cos(\tilde{\mathbf{r}}^m_n, \tilde{\mathbf{r}}^m_n)/\tau}}{\sum_{m'=1}^{M} e^{\cos(\tilde{\mathbf{r}}^m_n, \tilde{\mathbf{r}}^{m'}_n)/\tau}} \tag{9}$$

where $\tilde{\mathbf{r}}^m_n \in \mathbb{R}^d$ is linearly projected from $\mathbf{r}^m_n$ with normalization. By applying $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{Dis}}$, the extraction of modality-specific feature $\mathbf{h}^m_n$ mixed with unique and redundant features is constrained. With the optimizing of $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{Dis}}$, the unimodal uncertainty noise mixed in the modality-specific feature can be effectively removed.

4.2 Synergistic Distribution Alignment Module

In order to ensure that the synergistic component $\mathbf{s}^m_n$ captures the shared and consistent knowledge across modalities, Synergistic Distribution Alignment (SDA) module is proposed to perform the feature space alignment, which is illustrated in Figure 4.

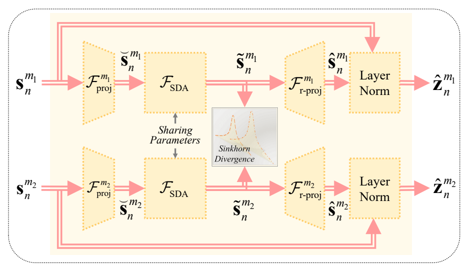


Figure 4: The illustration of Synergistic Distribution Alignment (SDA) module.

To achieve the alignment, SDA first linearly projects $\mathbf{s}^m_n \in \mathbb{R}^{d^m}$ to a subspace with d dimension shared by all modalities:

$$\breve{\mathbf{s}}^m_n = \mathcal{F}^m_{\text{proj}}(\mathbf{s}^m_n; \theta^m_{\text{proj}}) \in \mathbb{R}^d \quad \text{s.t. } m = \{1, \cdots, M\} \tag{10}$$

where θ^m_{proj} is the parameter of the m-th modality projection function $\mathcal{F}^m_{\text{proj}} : \mathbb{R}^{d^m} \to \mathbb{R}^d$. With all modalities in the same dimension, $\breve{\mathbf{s}}^m_n$ of each modality is able to propagate forward to the parameter-sharing MLP $\mathcal{F}_{\text{SDA}} : \mathbb{R}^d \to \mathbb{R}^d$ to achieve the space alignment:

$$\tilde{\mathbf{s}}^m_n = \mathcal{F}_{\text{SDA}}(\breve{\mathbf{s}}^m_n; \theta_{\text{SDA}}) \;\; \text{s.t. } m = \{1, \cdots, M\} \tag{11}$$

where θ_{SDA} is the parameter of MLP $\mathcal{F}_{\text{SDA}}$. To preserve the original model's structure and dimensionality, SDA linearly projects $\tilde{\mathbf{s}}^m_n$ back to the original space as:

$$\hat{\mathbf{s}}^m_n = \mathcal{F}^m_{\text{r-proj}}(\tilde{\mathbf{s}}^m_n; \theta^m_{\text{r-proj}}) \;\; \text{s.t. } m = \{1, \cdots, M\} \tag{12}$$

where $\theta^m_{\text{r-proj}}$ is the parameter of projection layer $\mathcal{F}^m_{\text{r-proj}} : \mathbb{R}^d \to \mathbb{R}^{d^m}$ in the m-th modality. To reduce the impact of back propagation on the gradient of model and the original semantic information shift[14; 33], we apply shortcut between $\mathbf{s}^m_n$ and $\hat{\mathbf{s}}^m_n$ with layer normalization:

$$\hat{\mathbf{z}}^m_n = \text{LayerNorm}(\hat{\mathbf{s}}^m_n + \mathbf{s}^m_n) \;\; \text{s.t. } m = \{1, \cdots, M\} \tag{13}$$

The key target of SDA module is to effectively constrain the extracted $\mathbf{s}^m_n$ contains the shared information among different modalities. However, the true distributions of $\mathbf{s}^m_n$ are unreachable. Although there have been works assuming them to be multivariate Gaussian distributions and trying to meet this assumption by applying Kullback-Leibler (KL) divergence [11; 19], [2] has pointed out the poor stability of KL-divergence for gradient descent. Inspired by [10; 26; 17], we employ the Sinkhorn-divergence defined in Equation (15), a variant of the optimal transport distance, to measure the difference between the distribution of latent embeddings from two modalities.

Definition 4.1 (Bi-modalities Batch Discrete Probability)**.** *Given a batch size N_c, the latent feature vectors in the m_1-th and m_2-th modalities are $\{\tilde{\mathbf{s}}_i^{m_1}\}_{i=1}^{N_c}$ and $\{\tilde{\mathbf{s}}_j^{m_2}\}_{j=1}^{N_c}$ respectively. Let Δ_{N_c} denotes the probability simplex of $\mathbb{R}^{N_c}$, and $\boldsymbol{a} = [a_1, \cdots, a_{N_c}] \in \Delta_{N_c}$, $\boldsymbol{b} = [b_1, \cdots, b_{N_c}] \in \Delta_{N_c}$. $\delta_{\tilde{\mathbf{s}}_i^*}$ refers to a point mass located at coordinate $\tilde{\mathbf{s}}_i^* \in \mathbb{R}^d$, We have the bi-modalities batch discrete probability:*

$$\alpha = \sum_{i=1}^{N_c} a_i \delta_{\tilde{\mathbf{s}}_i^{m_1}}, \quad \beta = \sum_{j=1}^{N_c} b_j \delta_{\tilde{\mathbf{s}}_j^{m_2}} \quad \textit{s.t.} \quad m_1, m_2 \in \{1, \cdots, M\} \ \textit{and} \ m_1 \neq m_2 \tag{14}$$

Based on Definition 4.1, according to [10; 26; 17], the Sinkhorn-divergence based term is defined:

$$\begin{aligned} \min \mathcal{L}_{\text{SDA}} &\overset{\text{def.}}{=} \frac{2}{M(M-1)} \sum_{\substack{m_1, m_2 \in \{1, \cdots, M\} \\ m_1 < m_2}} \mathcal{L}_{\text{div}}(m_1, m_2) \\ \text{s.t.} \ \ \mathcal{L}_{\text{div}}(m_1, m_2) &= \text{OT}(\alpha, \beta) - 0.5(\text{OT}(\alpha, \alpha) + \text{OT}(\beta, \beta)) \end{aligned} \tag{15}$$

Here, $\text{OT}(\cdot,\cdot)$ is the total optimal transport cost between two distributions solved by the regular optimal transport defined as $\text{OT}(\alpha, \beta) = \min_{\mathbf{T} \in \Pi(\alpha,\beta)} < \mathbf{T}, \mathbf{M} >_{\text{F}}$, where $\mathbf{M} \in \mathbb{R}^{N_c \times N_c}$ is the cost matrix with each element calculated with cosine similarity, *i.e.*, $\mathbf{M}_{i,j} = 1 - \cos(\tilde{s}_i^{m_1}, \tilde{s}_j^{m_2})$. $< \cdot, \cdot >_{\text{F}}$ is the Frobenius dot-product. The constrain $\Pi(\alpha, \beta) := \{\mathbf{T} \in \mathbb{R}_+^{N_c \times N_c} | \sum_i^{N_c} \mathbf{T}_{i,j} = b_j, \sum_j^{N_c} \mathbf{T}_{i,j} = a_i\}$ enforces $\mathbf{T}$ to have α, β as its marginals. $\mathbf{T}$ is the transport plan matrix to be optimized by Sinkhorn algorithm [7].

By minimizing Equation (15), the shared and consistent information across modalities can be retained. Finally, the two task-related components $\mathbf{u}_n^m$ and $\hat{\mathbf{z}}_n^m$ are aggregated by:

$$\tilde{\mathbf{z}}_n^m = \hat{\mathbf{z}}_n^m + \mathbf{u}_n^m \tag{16}$$

4.3 Training Loss Function

To sum up, the optimization target of our proposed FCD can be formed as:

$$\mathcal{L}_{\text{FCD}} = \lambda_1 \mathcal{L}_{\text{MI}} + \lambda_2 \mathcal{L}_{\text{Dis}} + \lambda_3 \mathcal{L}_{\text{SDA}} \tag{17}$$

By combining the original loss function (Equation (2)) and Equation (17), the total loss of an effective multimodal representation learning model with plug-and-play FCD without changing model structure is:

$$\mathcal{L}_{\text{overall}} = \mathcal{L} + \mathcal{L}_{\text{FCD}} = \underbrace{\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{reg}}}_{\text{Original Loss}} + \underbrace{\lambda_1 \mathcal{L}_{\text{MI}} + \lambda_2 \mathcal{L}_{\text{Dis}} + \lambda_3 \mathcal{L}_{\text{SDA}}}_{\text{FCD Loss}} \tag{18}$$

5 Experiments

5.1 Datasets and Evaluation Metrics

Datasets. We design and conduct our experiments on 5 widely used multimodal datasets, *i.e.*, CMU-MOSI [41], CMU-MOSEI [42], MSVA-Single [27], UPMC-Food101 [3], and HFM [4]. Please refer to Appendix Section C for more descriptions about these datasets.

Evaluation Metrics. Following [9], we report our experimental results on CMU-MOSI and CMU-MOSEI datasets with *the mean absolute error (MAE)*, *Pearson correlation (Corr)*, *binary classification accuracy (Acc-2)* and *weighted F1 score (F1)*. Following [11], we report the results on the remaining datasets with *accuracy (Acc)* and *weighted F1 score (F1)*. Please refer to Appendix Section C for more details about these metrics.

5.2 Experimental Settings and Implementation Details

Our experiments are conducted on 4 NVIDIA RTX 4090 24GB GPUs with PyTorch [29] framework. We control the system environment for all experiments to be consistent and reproduced all the base comparison experiments using the hyper-parameters reported in their original papers. Please refer to Appendix for more experiments, including sensitive analysis and computational overhead analysis.

Table 1: Quantitative results of SOTA methods. The left side of "/" in Acc-2 and F1 is computed as "negative/non-negative (non-exclude 0)" and the right side is computed as "negative/positive (exclude 0)". All results are scaled ($\times 100$). Better results in each compared method are highlighted in **bold**.

Datasets		CMU-MOSI				CMU-MOSEI			
Methods		MAE↓	Corr↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-2↑	F1↑
Self-MM	Base	72.51	79.44	82.80/84.30	82.78/84.33	54.04	75.65	83.26/85.20	83.40/84.98
	Ours	**68.29**	**80.62**	**84.95/87.09**	**84.93/87.11**	**52.49**	**76.98**	**85.19/86.08**	**85.12/85.77**
MMIM	Base	73.81	78.17	83.24/85.06	83.21/85.09	55.34	75.18	82.14/84.76	82.49/84.66
	Ours	**69.86**	**80.91**	**84.55/85.98**	**84.56/86.03**	**54.41**	**76.83**	**83.97/85.72**	**83.98/85.74**
MCL-MCF	Base	70.05	79.59	83.97/86.13	83.75/85.99	54.21	76.68	80.83/84.95	81.40/84.96
	Ours	**69.29**	**80.29**	**84.99/87.04**	**84.84/86.96**	**53.68**	**77.82**	**84.40/85.53**	**84.41/85.27**
AtCAF	Base	72.80	79.57	83.67/85.06	83.61/85.04	53.96	75.99	82.98/84.40	83.09/84.17
	Ours	**69.65**	**80.32**	**84.11/86.13**	**84.01/86.09**	**53.13**	**77.38**	**84.78/85.58**	**84.73/85.28**
MMML	Base	61.11	86.83	85.86/88.11	85.75/88.06	51.95	80.74	85.46/87.01	85.49/86.81
	Ours	**59.86**	**87.55**	**88.48/90.24**	**88.44/90.23**	**50.01**	**81.78**	**86.63/88.08**	**86.63/87.98**

The main hyper-parameters in this paper consist of two parts, *i.e.*, the hyper-parameters in each multi-modal intermediate fusion method (*e.g.*, learning rate, batch size) and the ones of FCD (λ_1, λ_2, λ_3). The hyper-parameters of FCD and their sensitive analysis are summarized in Appendix Section E, and please refer to the original papers for other hyper-parameters of each base model. Moreover, we report the computational overhead and complexity analysis in Appendix Section F. We fix the temperature parameter $\tau = 0.07$ in InfoNCE loss for the most common cases.

5.3 Quantitative Studies

To evaluate the effectiveness of our proposed FCD, we select 5 recently proposed state-of-the-art (SOTA) multimodal sentiment analysis methods as our evaluation baseline on CMU-MOSI and CMU-MOSEI datasets, *i.e.*, Self-MM [39], MMIM [13], MCL-MCF [9], AtCAF [15], and MMML [37]. And 4 multimodal fusion methods on the remaining 3 datasets are compared: MMBT [18], CLMLF [20], MVCN [36], and URMF [11]. All these methods are intermediate fusion methods and please refer to Appendix Section D for more details.

The results on CMU-MOSI and CMU-MOSEI datasets are shown in Table 1, and the results on the other 3 datasets are reported in Table 2. The *Base* experiments are conducted on the corresponding datasets using the official implementations and the hyper-parameters reported in their papers. Then we add FCD to each method and reproduce them again with the original hyper-parameters fixed, which are represented by "*Ours*". By doing so, we are able to control the uniqueness that affects the experimental results and improve their credibility.

Table 2: Quantitative results of SOTA methods on MVSA-Single (MVSA-S), UPMC-Food101 (Food101), and HFM datasets. All results are scaled ($\times 100$). Better results in each compared method are highlighted in **bold**.

Datasets		MVSA-S		Food101		HFM	
Methods		Acc	F1	Acc	F1	Acc	F1
MMBT	Base	74.76	74.76	91.22	91.22	-	-
	Ours	**76.30**	**76.30**	**91.72**	**91.72**	-	-
CLMLF	Base	70.67	69.03	-	-	84.77	84.76
	Ours	**72.22**	**71.62**	-	-	**85.55**	**85.52**
MVCN	Base	72.44	71.25	-	-	84.89	84.90
	Ours	**75.33**	**74.78**	-	-	**85.31**	**85.32**
URMF	Base	72.25	71.11	92.27	92.26	-	-
	Ours	**74.57**	**74.22**	**92.94**	**92.96**	-	-

From Table 1 and Table 2, we can find that our proposed FCD is able to improve the performance among various evaluation metrics on all tested regression tasks and classification tasks. This indicates that the heterogeneity and aleatoric uncertainty of the multimodal data should be considered simultaneously to generate valid multimodal representation. Specifically, MMML achieves the best performances on both CMU-MOSI and CMU-MOSEI datasets after applying FCD, which may be caused by the superior behavior and complexity of MMML. It has the longest training overhead per epoch (see Appendix Section F). On the other hand, MMBT has the highest *Acc* and *F1* on MVSA-S dataset. URMF performs better than MMBT on Food101 dataset on both *Acc* and *F1* metrics. CLMLF shows the better performance than MVCN on HFM dataset on both tested metrics. The improvements of methods are various. This may be caused by the different unimodal encoding approaches, which directly determines the quality and semantic richness of unimodal features and thereafter affect the effectiveness of FCD.

There exists a significant difference with other methods that Self-MM considered both consistent and complementary information in multimodal data. It took the heterogeneity of multimodal feature space into consideration and proposed to shift annotations in their unimodal feature space. This proves that the consistent and complementary information within multimodal data should be effectively excavated at the same time, which also coincides with our idea.

Besides, compared with other methods, the *Corr* of MCL-MCF with FCD has less improvement on both CMU-MOSI and CMU-MOSEI datasets. This may be caused by the assumption of MCL-MCF where the fusion is considered as a progressive process. In this situation, the consistent information plays the dominant role and the aggregated complementary information is gradually weakened as the fusion process progresses.

5.4 Ablation Studies

To evaluate each component used in FCD, ablation studies are conducted. Specifically, we conduct ablation experiments on term $\mathcal{L}_{\text{MI}}$, $\mathcal{L}_{\text{Dis}}$ and $\mathcal{L}_{\text{SDA}}$ by setting $\lambda_1 = 0$, $\lambda_2 = 0$ and $\lambda_3 = 0$, respectively. Besides, we also evaluate the effectiveness of involving unique component $\mathbf{u}_n^m$ in Equation (16), where only synergistic feature $\hat{\mathbf{z}}_n^m$ is used for fusion. And the shortcut Equation (13), where $\hat{\mathbf{s}}_n^m$ is directly fused with $\mathbf{u}_n^m$. The results can be seen in Table 3 and the results of *Base* and *Ours* are also shown in the first line and the last line, respectively.

Table 3: Ablation studies results on CMU-MOSI dataset with Self-MM. The ablation terms are marked as "✓" if they are not ablated, otherwise they are marked as "-".

$\mathcal{L}_{\text{MI}}$	$\mathcal{L}_{\text{Dis}}$	$\mathcal{L}_{\text{SDA}}$	Eq. (16)	Eq. (13)	MAE	Corr	Acc-2	F1
		Base			72.51	79.44	82.80/84.30	82.78/84.33
-	✓	✓	✓	✓	70.76	79.65	82.94/84.91	82.85/84.88
✓	-	✓	✓	✓	68.83	80.21	84.42/86.53	84.39/86.38
✓	✓	-	✓	✓	68.66	80.47	83.38/85.98	83.15/85.84
✓	✓	✓	-	✓	69.45	80.33	83.24/84.91	83.19/84.91
✓	✓	✓	✓	-	68.92	80.17	84.02/86.13	83.94/86.07
✓	✓	✓	✓	✓	**68.29**	**80.62**	**84.95/87.09**	**84.93/87.11**

When $\lambda_1 = 0$, the mutual information loss between $\mathbf{u}_n^m$ and y_n would not be added to the total loss function. In other words, the backdoor-adjustment is not applied to effectively decompose the redundant feature $\mathbf{r}_n^m$ and the unique feature $\mathbf{u}_n^m$. The unimodal uncertainty noise may be still mixed in the complementary information. Meanwhile, since the unique feature is modality-specific, the task-related information may be also mixed in $\mathbf{r}_n^m$, resulting in the worst performance in all ablation cases. This case proves that CCD can extract task-related information from the mixed features.

When $\lambda_2 = 0$, the discriminative information between $\mathbf{r}_n^m$ is neglected. In this case, $\mathbf{h}_n^m$ extracted by $\mathcal{F}_{\text{h}}$ may be not modality-specific. In this case, the complementary information may not be fully exploited, resulting in task-related information loss.

When $\lambda_3 = 0$, there is no constrains on the synergistic information alignment. It means there is no supervision on the aligned feature space, which means the features are probably not modality-invariant and contains modality-specific information.

When removing the unique component $\mathbf{u}_n^m$ Equation (16), *i.e.*, $\tilde{\mathbf{z}}_n^m = \hat{\mathbf{z}}_n^m$, compared with the full FCD, the performance is worse due to the complementary information loss. Therefore, the unique component is significant for multimodal representation learning.

When the shortcut between $\mathbf{s}_n^m$ and $\hat{\mathbf{s}}_n^m$ is removed, the gradient information and original semantic knowledge may be lost. Thus, the shortcut path is able to retain the original information.

5.5 Case Study

To validate the corresponding parts of task-related components $\mathbf{s}_n^m$ and $\mathbf{u}_n^m$ in the original data and the interpretability of FCD, we employ Grad-CAM [30] to visualize the decomposition results of CLMLF on MVSA-S dataset. The results is shown in Figure 5.

From Figure 5, we can find that there is a correlated semantic similarity shared by the synergistic components of different modalities (*e.g.*, building collapse). On the other hand, the scattered bricks on the ground from visual modality and "fire" from textual modality provide the unique information, which is related to the annotation *Negative*. This indicates that FCD is able to capture the synergistic and unique information. Besides, for both modalities, there may exist importance overlap between synergistic and unique information, such as the word "building" is both more important in textual modality and both more attentions are paid on the buildings in the distance of the visual modality. For synergistic information, it focuses on the semantic consistency to construct the semantic correspondence between different modalities. For unique information, it focuses on the unimodal contextual information learned from the whole dataset. Taking the "building" in textual modality as an example, it may experience more concurrent together with negative words like "disasters" or "collapse". This also proves the necessity of simultaneously excavating the consistency and complementary information.

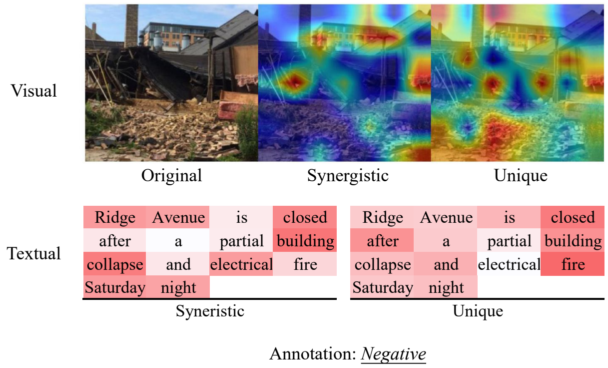


Figure 5: The Grad-CAM of task-related components $\mathbf{s}_n^m$ and $\mathbf{u}_n^m$ extracted by CLMLF trained on MVSA-S. The upper part of the figure is the original, synergistic and unique data of visual modality. And the lower half of the figure is the synergistic and unique data of textual modality.

5.6 Sensitivity of Batch Size

Since the $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{Dis}}$ loss functions used in the FCD are essentially contrastive learning, whose effectiveness may be affected by the number of negative samples provided by the batch size, we validate the sensitivity of batch size. We conduct this experiment on Self-MM equipped with FCD on CMU-MOSI dataset. Practically, the batch size doubles from 8 to 128.

Table 4: The sensitivity of batch size on Self-MM (with FCD) on CMU-MOSI dataset.

Batch Size	MAE	Corr	Acc-2	F1
8	67.83	80.34	84.74/86.96	84.74/87.01
16	68.02	79.86	84.43/86.27	84.48/86.31
32	68.29	80.62	84.95/87.09	84.93/87.11
64	68.11	80.27	85.15/87.41	85.16/87.42
128	67.94	80.57	85.74/87.85	85.71/87.83

The results are shown in Table 4. The results show that as the batch size increases, the model achieves consistently better performance across various metrics. One possible reason is that a larger batch provides a wider range of negative samples for the contrastive loss, which makes the optimization of $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{Dis}}$ more effective. With more diverse negatives, the model can better separate similar but distinct representations, leading to stronger feature discrimination and better overall performance.

6 Conclusion

In this paper, we propose a plug-and-play module, Feature Causality Decomposition (FCD), to solve the existence of heterogeneity and aleatoric uncertainty within multimodal data by decomposing the unimodal feature into its synergistic, unique and redundant components from causality perspective. Based on whether it can be aligned with other modalities, FCD first uses Causality Components Decomposition (CCD) module to disentangle the unimodal feature into two parts: modality-specific and modality-invariant components, which contains the synergistic information shared by various heterogeneous modalities. Then backdoor-adjustment is applied to remove the redundant information and retain the task-related component in the modality-specific part, it is optimized by maximizing the mutual information between the unique component and the annotation. Besides, the Sinkhorn divergence is employed to narrow the difference of synergistic embeddings among modalities. Extensive experiments prove the effectiveness of FCD.

Acknowledgements

This work is partly supported by the National Science and Technology Major Project (2024YFF1206600), the National Natural Science Foundation of China (62306118, 62325204, U21A20520), the Fundamental Research Funds for the Central Universities (2025ZYGXZR054), China University Industry-University-Research Innovation Fund (2023RY031).

References

[1] Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=HyxQzBceg 3

[2] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017) 6

[3] Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13. pp. 446–461. Springer (2014) 7

[4] Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 2506–2515 (2019) 7

[5] Cao, B., Xia, Y., Ding, Y., Zhang, C., Hu, Q.: Predictive dynamic fusion. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=LYpGLrC4oq 2, 27

[6] Chen, Z., Hu, L., Li, W., Shao, Y., Nie, L.: Causal intervention and counterfactual reasoning for multi-modal fake news detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 627–638 (2023) 3

[7] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013) 7

[8] Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O.: Simmmdg: A simple and effective framework for multi-modal domain generalization. Advances in Neural Information Processing Systems **36**, 78674–78695 (2023) 3

[9] Fan, C., Zhu, K., Tao, J., Yi, G., Xue, J., Lv, Z.: Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis. IEEE Transactions on Affective Computing (2024) 1, 2, 3, 4, 5, 7, 8, 23, 24

[10] Feydy, J., Séjourné, T., Vialard, F.X., Amari, S.i., Trouvé, A., Peyré, G.: Interpolating between optimal transport and mmd using sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 2681–2690. PMLR (2019) 2, 6, 7

[11] Gao, Z., Jiang, X., Xu, X., Shen, F., Li, Y., Shen, H.T.: Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26876–26885 (2024) 1, 2, 3, 6, 7, 8, 23, 24, 25

[12] Guan, W., Wen, H., Song, X., Yeh, C.H., Chang, X., Nie, L.: Multimodal compatibility modeling via exploring the consistent and complementary correlations. In: Proceedings of the 29th ACM international conference on multimedia. pp. 2299–2307 (2021) 1

[13] Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 9180–9192. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.723, https://aclanthology.org/2021.emnlp-main.723/ 1, 2, 3, 8, 23, 24

[14] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 6

[15] Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., Huang, X.: Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. Information Fusion **114**, 102725 (2025) 1, 3, 5, 8, 23, 24

[16] Jaynes, E.T.: Probability theory: The logic of science. Cambridge university press (2003) 2, 5, 21

[17] Ji, Z., Tian, X., Liu, Y.: Affakt: A hierarchical optimal transport based method for affective facial knowledge transfer in video deception detection. Proceedings of the AAAI Conference on Artificial Intelligence **39**(2), 1336–1344 (Apr 2025). https://doi.org/10.1609/aaai.v39i2.32123, https://ojs.aaai.org/index.php/AAAI/article/view/32123 6, 7

[18] Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950 (2019) 1, 8, 24

[19] Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022), https://arxiv.org/abs/1312.6114 6

[20] Li, Z., Xu, B., Zhu, C., Zhao, T.: CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Findings of the Association for Computational Linguistics: NAACL 2022. pp. 2282–2294. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.175, https://aclanthology.org/2022.findings-naacl.175/ 1, 3, 8, 24

[21] Lin, R., Hu, H.: Multi-task momentum distillation for multimodal sentiment analysis. IEEE Transactions on Affective Computing (2023) 5

[22] Liu, N., Wang, X., Wu, L., Chen, Y., Guo, X., Shi, C.: Compact graph structure learning via mutual information compression. In: Proceedings of the ACM web conference 2022. pp. 1601–1610 (2022) 5

[23] Liu, Y., Li, G., Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 11624–11641 (2023) 3, 5

[24] Martínez-Sánchez, Á., Arranz, G., Lozano-Durán, A.: Decomposing causality into its synergistic, unique, and redundant components. Nature Communications **15**(1), 9296 (2024) 2

[25] Neuberg, L.G.: Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. Econometric Theory **19**(4), 675–685 (2003) 2, 3, 5, 22

[26] Nguyen, T.T., Luu, A.T.: Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 11103–11111 (2022) 6, 7

[27] Niu, T., Zhu, S., Pang, L., El Saddik, A.: Sentiment analysis on multi-view social data. In: MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22. pp. 15–27. Springer (2016) 7, 24

[28] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3, 5, 6

[29] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019) 7

[30] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 9

[31] Tian, H., Tao, Y., Pouyanfar, S., Chen, S.C., Shyu, M.L.: Multimodal deep representation learning for video classification. World Wide Web **22**, 1325–1341 (2019) 3

[32] Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting. vol. 2019, p. 6558. NIH Public Access (2019) 24

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6

[34] Walters, P.: An introduction to ergodic theory, vol. 79. Springer Science & Business Media (2000) 23

[35] Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15878–15887 (June 2023) 3

[36] Wei, Y., Yuan, S., Yang, R., Shen, L., Li, Z., Wang, L., Chen, M.: Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5240–5252 (2023) 1, 8, 25

[37] Wu, Z., Gong, Z., Koo, J., Hirschberg, J.: Multimodal multi-loss fusion network for sentiment analysis. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3588–3602. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.197, https://aclanthology.org/2024.naacl-long.197/ 8, 24

[38] Yan, J., Deng, C., Huang, H., Liu, W.: Causality-invariant interactive mining for cross-modal similarity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 2, 3

[39] Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 10790–10797 (2021) 1, 8, 24

[40] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 (2017) 24

[41] Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016) 7

[42] Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246 (2018) 7

[43] Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J.T., Peng, X.: Provable dynamic fusion for low-quality multimodal data. In: International conference on machine learning. pp. 41753–41769. PMLR (2023) 23, 27

[44] Zheng, X., Lyu, Y., Wang, L.: Learning modality-agnostic representation for semantic segmentation from any modalities. In: European Conference on Computer Vision. pp. 146–165. Springer (2024) 2, 3

[45] Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., Yu, L.: Triple disentangled representation learning for multimodal affective analysis. Information Fusion **114**, 102663 (2025) 2, 3, 4, 5

[46] Zou, X., Yan, Y., Xue, J.H., Chen, S., Wang, H.: Learn-to-decompose: cascaded decomposition network for cross-domain few-shot facial expression recognition. In: European Conference on Computer Vision. pp. 683–700. Springer (2022) 5

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims are made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses the limitations.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Proofs are in appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We will make the code public after acceptance.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We will make the code public after acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental details are given.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The experimental results don't contain error bars etc.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We give the computational platform.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We have discussed in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: We don't have these contents.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We have cited the works in our paper.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: We don't release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: We don't provide new assets with human.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

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

A Proof of Theorem 4.1

We first give the following lemma, which describes the probability distribution between two variables X and Y.

Lemma A.1. *For two variables X, Y defined on the measurable sets $\mathcal{X}$, $\mathcal{Y}$ with the same measure, and the probability distributions are $P(X)$, $P(Y)$. If $Y = \mathcal{F}(X)$, where $\mathcal{F}$ is a measure-preserving bijective funcion, we have $P(X) = P(Y)$.*

Proof. Since $\mathcal{F}$ is a measure-preserving bijective function, we have:

$$P(X \in \mathcal{X}) = P(\mathcal{F}^{-1}(Y) \in \mathcal{X}) = P(Y \in \mathcal{F}(\mathcal{X})) = P(Y \in \mathcal{Y}) \tag{19}$$

□

Corollary A.1. *For two variables X, Y from measurable sets $\mathcal{X}$, $\mathcal{Y}$ with their measures $\mu(\mathcal{X})$ and $\mu(\mathcal{Y})$. If $Y = \mathcal{F}(X)$ and $\mathcal{F}$ is a measure-preserving bijective function, there exists:*

$$P(X, Y) = P(X)\delta(Y - \mathcal{F}(X)) = P(Y)\delta(Y - \mathcal{F}(X)) \tag{20}$$

when $\mu(\mathcal{X}) = \mu(\mathcal{Y})$.

where $\delta(t)$ is the standard Dirac delta function which has the following properties for $t \in \mathbb{R}$:

1. $\delta(t) = 0$ for all $t \neq 0$
2. $\int_{-\infty}^{+\infty} \delta(t)\mathrm{d}t = 1$.

Proof. Based on Bayes' theorem, we have:

$$P(X, Y) = P(X)P(Y|X) \tag{21}$$

Since $\mathcal{F}$ is a bijective function, Y is uniquely determined by X, according to [16], $P(Y|X) = \delta(Y - \mathcal{F}(X))$. Based on Lemma A.1, we have:

$$P(X, Y) = P(X)P(Y|X) = P(X)\delta(Y - \mathcal{F}(X)) = P(Y)\delta(Y - \mathcal{F}(X)) \tag{22}$$

□

Corollary A.2. *For three variables X, Y and Z defined on the measurable sets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, and the joint probability distributions are $P(X, Z)$, $P(Y, Z)$. If $Y = \mathcal{F}(X)$, where $\mathcal{F}$ is a measure-preserving bijective function, X is independent with Z, and Y is also independent with Z, we have $P(X, Z) = P(Y, Z)$.*

Proof. Let $\mu(\mathcal{X})$ is the measure of $\mathcal{X}$. Since $Y = \mathcal{F}(X)$, where $\mathcal{F}$ is a measure-preserving bijective function, then $\mathcal{X}$ and $\mathcal{Y}$ have the same measure, i.e., $\mu(\mathcal{X}) = \mu(\mathcal{Y})$. Besides, the measures of the joint domains $\mathcal{X} \times \mathcal{Z}$, $\mathcal{Y} \times \mathcal{Z}$ can be formed as $\mu(\mathcal{X} \times \mathcal{Z})$ and $\mu(\mathcal{Y} \times \mathcal{Z})$, respectively. Based on the Fubini's Theorem, we have:

$$\mu(\mathcal{X} \times \mathcal{Z}) = \mu(\mathcal{X})\mu(\mathcal{Z}) = \mu(\mathcal{Y})\mu(\mathcal{Z}) = \mu(\mathcal{Y} \times \mathcal{Z}) \tag{23}$$

Thus, the measures of joint domains $\mathcal{X} \times \mathcal{Z}$ and $\mathcal{Y} \times \mathcal{Z}$ are the same. Since $\mathcal{F}$ is a measure-preserving bijective function, X is independent with Z, and Y is also independent with Z, based on Lemma A.1, we have:

$$\begin{aligned} &P((X, Z) \in (\mathcal{X} \times \mathcal{Z})) = P(X \in \mathcal{X})P(Z) \\ =\ &P(\mathcal{F}^{-1}(Y) \in \mathcal{X})P(Z) = P(Y \in \mathcal{F}(\mathcal{X}))P(Z) = P((Y, Z) \in (\mathcal{Y} \times \mathcal{Z})) \end{aligned} \tag{24}$$

□

Corollary A.3. *For three variables X, Y, Z from measurable sets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, where X is independent with Z and Y is also independent with Z. If $Y = \mathcal{F}(X)$ and $\mathcal{F}$ is a measure-preserving bijective function, there exists:*

$$P(X, Y, Z) = P(X, Z)\delta(Y - \mathcal{F}(X)) = P(Y, Z)\delta(Y - \mathcal{F}(X)) \tag{25}$$

when $\mu(\mathcal{X}) = \mu(\mathcal{Y})$.

Proof. Based on Bayes' theorem, we have:

$$P(X, Y, Z) = P(X, Z)P(Y|X, Z) \tag{26}$$

Since $\mathcal{F}$ is a bijective function, Y is uniquely determined by X. Thus, X is independent with Z and Y is also independent with Z, then we have:

$$P(Y|X, Z) = \frac{P(Y, X, Z)}{P(X, Z)} = \frac{P(Y, X)P(Z)}{P(X)P(Z)} = \frac{P(Y, X)}{P(X)} = P(Y|X)$$

According to Corollary A.1 and Corollary A.2, and $P(Y|X)$ is a Dirac delta function $\delta(Y - \mathcal{F}(X))$, we have:

$$\begin{aligned} P(X, Y, Z) &= P(X, Z)P(Y|X, Z) \\ &= P(X, Z)P(Y|X) = P(X, Z)\delta(Y - \mathcal{F}(X)) = P(Y, Z)\delta(Y - \mathcal{F}(X)) \end{aligned} \tag{27}$$

□

Based on Corollary A.1 and Corollary A.3, we can prove the Theorem 4.1 as follows.

Proof. For the hidden features $\mathbf{h}_n^m$, $\mathbf{u}_n^m$ and the ground truth y_n, we are seeking to maximizing the expectation of the conditional mutual information between $\mathbf{u}_n^m$ and the ground truth y_n, when $\mathbf{h}_n^m$ is given. From the probability perspective, the $\mathbf{h}_n^m$ and $\mathbf{u}_n^m$ are both independent with the given ground truth y_n. Based on the definition of conditional mutual information and the Bayes' theorem, we have:

$$\begin{aligned} &\max \mathbb{E}_{P(\mathbf{h}_n^m)}\left[I(\mathbf{u}_n^m; y_n|\mathbf{h}_n^m)\right] = \max \mathbb{E}_{P(\mathbf{h}_n^m)}\left[\mathbb{E}_{P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m)}\left[\log \frac{P(\mathbf{u}_n^m, y_n|\mathbf{h}_n^m)}{P(\mathbf{u}_n^m|\mathbf{h}_n^m)P(y_n|\mathbf{h}_n^m)}\right]\right] \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(\mathbf{h}_n^m)P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n|\mathbf{h}_n^m)}{P(\mathbf{u}_n^m|\mathbf{h}_n^m)P(y_n|\mathbf{h}_n^m)} \mathrm{d}y_n \mathrm{d}\mathbf{h}_n^m \mathrm{d}\mathbf{u}_n^m \mathrm{d}\mathbf{h}_n^m \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(\mathbf{h}_n^m)P(y_n|\mathbf{u}_n^m, \mathbf{h}_n^m)P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n|\mathbf{h}_n^m)}{P(\mathbf{u}_n^m|\mathbf{h}_n^m)P(y_n|\mathbf{h}_n^m)} \mathrm{d}y_n \mathrm{d}\mathbf{h}_n^m \mathrm{d}\mathbf{u}_n^m \mathrm{d}\mathbf{h}_n^m \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(\mathbf{h}_n^m)P(y_n|\mathbf{u}_n^m, \mathbf{h}_n^m)P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n|\mathbf{h}_n^m)P(\mathbf{h}_n^m)}{P(\mathbf{u}_n^m|\mathbf{h}_n^m)P(y_n|\mathbf{h}_n^m)P(\mathbf{h}_n^m)} \mathrm{d}y_n \mathrm{d}\mathbf{h}_n^m \mathrm{d}\mathbf{u}_n^m \mathrm{d}\mathbf{h}_n^m \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(\mathbf{h}_n^m)P(y_n|\mathbf{u}_n^m, \mathbf{h}_n^m)P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m)}{P(\mathbf{u}_n^m, \mathbf{h}_n^m)P(y_n|\mathbf{h}_n^m)} \mathrm{d}y_n \mathrm{d}\mathbf{h}_n^m \mathrm{d}\mathbf{u}_n^m \mathrm{d}\mathbf{h}_n^m \end{aligned} \tag{28}$$

Note that the distribution of the ground truth $P(y_n)$ doesn't change with $\mathbf{h}_n^m$, thus we have $P(y_n|\mathbf{h}_n^m) = P(y_n)$. Besides, based on the core idea of backdoor-adjustment that causal intervention is applied for the confounder to cut off the causal relation between the treatment and the confounder [25], we apply the causal intervention on $\mathbf{h}_n^m$ to turn it into its counterfactual form $\mathbf{h}_n^{m\prime}$, and cut off the causal relation between $\mathbf{h}_n^m$ and $\mathbf{u}_n^m$. From Corollary A.1 and Corollary A.3, the Equation (28) can be further transformed to Equation (29).

$$\begin{aligned} &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(\mathbf{h}_n^m)P(y_n|\mathbf{u}_n^m, \mathbf{h}_n^m)P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n)\delta(\mathbf{u}_n^m - \mathcal{F}_{\text{mb}}^m(\mathbf{h}_n^m; \theta_{\text{mb}}^m))}{P(\mathbf{u}_n^m)\delta(\mathbf{u}_n^m - \mathcal{F}_{\text{mb}}^m(\mathbf{h}_n^m; \theta_{\text{mb}}^m))P(y_n)} \mathrm{d}y_n \mathrm{d}\mathbf{h}_n^m \mathrm{d}\mathbf{u}_n^m \mathrm{d}\mathbf{h}_n^m \\ &\xrightarrow[\text{Intervention}]{\text{Causal}} \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} \underbrace{P(\mathbf{h}_n^{m\prime})P(y_n|\mathbf{u}_n^m, \mathbf{h}_n^{m\prime})}_{\text{Backdoor-adjustment}} P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n)}{P(\mathbf{u}_n^m)P(y_n)} \mathrm{d}y_n \mathrm{d}\mathbf{h}_n^m \mathrm{d}do(\mathbf{u}_n^m) \mathrm{d}\mathbf{h}_n^{m\prime} \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} \int_{\mathbb{R}^{d^m}} \underbrace{P(\mathbf{h}_n^{m\prime})P(y_n|\mathbf{u}_n^m, \mathbf{h}_n^{m\prime})}_{\text{Backdoor-adjustment}} \mathrm{d}\mathbf{h}_n^{m\prime} \int_{\mathbb{R}^{d^m}} P(\mathbf{u}_n^m, \mathbf{h}_n^m) \mathrm{d}\mathbf{h}_n^m \log \frac{P(\mathbf{u}_n^m, y_n)}{P(\mathbf{u}_n^m)P(y_n)} \mathrm{d}y_n \mathrm{d}\mathbf{u}_n^m \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(y_n|do(\mathbf{u}_n^m))P(do(\mathbf{u}_n^m)) \log \frac{P(do(\mathbf{u}_n^m), y_n)}{P(do(\mathbf{u}_n^m))P(y_n)} \mathrm{d}y_n \mathrm{d}do(\mathbf{u}_n^m) \\ &= \max \int_{\mathbb{R}^{d^m}} \int_{\mathbb{R}} P(do(\mathbf{u}_n^m), y_n) \log \frac{P(do(\mathbf{u}_n^m), y_n)}{P(do(\mathbf{u}_n^m))P(y_n)} \mathrm{d}y_n \mathrm{d}do(\mathbf{u}_n^m) \\ &= \max I(do(\mathbf{u}_n^m); y_n). \end{aligned} \tag{29}$$

where $\mathcal{F}_{\text{mb}}^m(\cdot; \theta_{\text{mb}}^m)$ is the bijective decomposition function with parameter θ_{mb}^m (the URD module in our paper). $do(\cdot)$ is *do*-operator for causal intervention in backdoor-adjustment. $\mathbf{h}_n^{m\prime}$ is the $\mathbf{h}_n^m$ after causal intervention. From Equation (28) and Equation (29), Theorem 4.1 holds. □

B Pseudo Code for Training Process

We first give the training process to show the reversibility of $\mathcal{F}^m_{\text{URD-d}}$ and $\mathcal{F}^m_{\text{URD-f}}$.

Algorithm 1: Enforce Full Rank in PyTorch style

Input: Multimodal model with FCD Model

for $m \leftarrow 1$ **to** M **do**
  `// Get parameters of linear layers in URD`
  $\theta^m_{\text{URD-d}}, \theta^m_{\text{URD-f}} \leftarrow$ `get` (Model.CCD.URD, Linear);
  `// SVD`
  $U^m_{\text{URD-d}}, \Sigma^m_{\text{URD-d}}, V^m_{\text{URD-d}} \leftarrow$ `SVD` ($\theta^m_{\text{URD-d}}$.weight);
  $U^m_{\text{URD-f}}, \Sigma^m_{\text{URD-f}}, V^m_{\text{URD-f}} \leftarrow$ `SVD` ($\theta^m_{\text{URD-f}}$.weight);
  `// Enforce Full Rank (Invertible)`
  $\Sigma^m_{\text{URD-d}} \leftarrow$ `torch.maximum` $(\Sigma^m_{\text{URD-d}}, 1e^{-5})$;
  $\Sigma^m_{\text{URD-f}} \leftarrow$ `torch.maximum` $(\Sigma^m_{\text{URD-f}}, 1e^{-5})$;
  `// Reconstruct`
  $\theta^m_{\text{URD-d}}$.weight $\leftarrow$ `nn.Parameter` $\left(U^m_{\text{URD-d}}\Sigma^m_{\text{URD-d}}V^{m\top}_{\text{URD-d}}\right)$;
  $\theta^m_{\text{URD-f}}$.weight $\leftarrow$ `nn.Parameter` $\left(U^m_{\text{URD-f}}\Sigma^m_{\text{URD-f}}V^{m\top}_{\text{URD-f}}\right)$;
end

Algorithm 2: Training process of our process method in PyTorch style

Input: Multimodal dataset $\mathcal{D}$, Multimodal model with FCD Model, Optimizer opt, Number of epochs N_e.

for e in $\{1, \cdots, N_e\}$ **do**
  $\mathcal{L}_{\text{overall}} \leftarrow$ Model.`forward`($\mathcal{D}$);
  $\mathcal{L}_{\text{overall}}$.`backward` ();
  opt.`step` ();
  Apply Algorithm 1 with model as input;
end

Besides, to achieve measure-preserving, we apply the following constraint on the weight matrix of $\mathcal{F}^m_{\text{URD-d}}$ and $\mathcal{F}^m_{\text{URD-f}}$. Note that the nesting of two measure-preserving functions is still measure-preserving [34].

Algorithm 3: Measure-preserving Linear Layer Forward

Input: Feature vector $\mathbf{x}$
Output: Forward output vector $\mathbf{x}'$
`// Get weight matrix and bias`
$\mathbf{W} \leftarrow \theta^m_*$.weight;
$\mathbf{b} \leftarrow \theta^m_*$.bias;
`// Construct skew symmetric matrix`
$\mathbf{A} \leftarrow \mathbf{W} - \mathbf{W}^\top$;
$\mathbf{W}' \leftarrow$ `torch.matrix_exp` $(\mathbf{A})$;
`// Linear forward`
$\mathbf{x}' \leftarrow \mathbf{x}\mathbf{W}'^\top + \mathbf{b}$;

C Introduction to Datasets and Evaluation Metrics in Our Experiments

CMU-MOSI and CMU-MOSEI are popularly used in Multimodal Semantic Analysis task [9; 13] with 3 modalities. Each sample from both of these datasets is annotated with a sentiment value ranging from -3 (strongly negative) to $+3$ (strongly positive), indicating the polarity and relative strength of the expressed sentiment within each sample. The former one contains 2199 utterance video segments taken from 93 YouTube, and the latter one contains 22,856 utterance video segments [11]. Table 5 summarizes the training, validation and test subset splits following [15; 43].

Table 5: Datasets splits (train, validation (val), and test) in our experiments.

DATASET	TRAIN	VAL	TEST	OVERALL
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
MVSA-SINGLE	1555	518	519	2592
UPMC FOOD101	62971	5000	22715	90686
HFM	19816	2410	2409	24635

MSVA-Single, UPMC Food101, and HFM datasets only have two modalities. MVSA-Single is a commonly used text-image sentiment dataset collected from Twitter. It has 3 categories: positive, neutral and negative with 1398, 724 and 470 samples, respectively [27]. UPMC Food101 dataset is usually used for multimodal image classification [11], which has 101 categories with about 100k images. HFM has two categories, *i.e.*, positive and negative.

We report our results with *the mean absolute error (MAE)*, which is calculated by averaging the absolute value between the predicted and ground truth; *Pearson correlation (Corr)* quantifies the extent to which predictions deviate from a linear relationship; *binary classification accuracy (Acc-2)* and *weighted F1 scores (F1)* are computed for both the negative/non-negative (non-exclude 0) [40] and negative/positive (exclude 0) [32], which is achieved by filtering our the samples whose annotation is 0. Following [11], we report the results on the remaining datasets with *accuracy (Acc)* and *weighted F1 score (F1)*.

D Introduction to Quantitative Experiment Methods

Due to the page limit, we introduce the methods incorporated in our quantitative experiment here.

- Self-MM [39] focused on restricted ability in capturing differentiated information within each modality due to the unified multimodal annotation, and proposed a self-supervised multi-task learning method to generate unimodal annotation. Besides, Self-MM shifts the generated annotation according to the relative distance to the class center in each modality space.
- MMIM [13] designed an MI based method to preserve critical task-related information that flows from the original input to the fusion representation. It employed a tight lower bound of MI and estimated the lower bound via likelihood maximization and Gaussian Mixture Model (GMM). Then a Contrastive Predictive Coding (CPC) loss was employed to retain the modality-invariant information in fused represetation.
- MCL-MCF [9] considered that fusion is a progressive process, and provided a hierarchical structure to maximize the maintenance of semantic information during different fusion level via contrastive learning. Besides, MCL-MCF used 1-D convolutional layers to fuse features at different level.
- AtCAF [15] started from casual inference perspective. It blocked the back-door path between text modality and target annotation via front-door adjustment. Then it applied counterfactual reasoning to the attention matrix integrated in cross-attention fusion process to improve the fusion robustness.
- MMML [37] proposed a Multimodal Multi-Loss Fusion Network that integrates pretrained audio and text encoders, cross- and self-attention mechanisms, and multi-loss training to enhance sentiment analysis. The model achieved state-of-the-art performance on various datasets, demonstrating the effectiveness of multimodal fusion and contextual modeling.
- MMBT [18] proposed a multimodal fusion method for text-image classification based on bitransformer. It jointly finetuned the pretrained unimodal encoders by mapping image embeddings to textual token space.
- CLMLF [20] focused on the token level multimodal fusion and employ contrastive learning to align the representations from different modalities.

- MVCN [36] focused on the challenge of modality heterogeneity in multimodal tasks, and proposed to filter redundant visual features based on sparsemax mechanism. Besides, it calibrated feature shift in representation space by minimizing the intra-class discrepancy.
- URMF [11] adopted a multivariate Gaussian distribution to represent spotty semantic instances in a noisy latent space and tried to eliminate the impact of unimodal aleatoric uncertainty to perform robust multimodal fusion via estimiting the Gaussian distribution behind features.

E Hyper-parameter Setting and Sensitive Analysis

The hyper-parameters setting of FCD used in different base methods are reported in Table 6. Since FCD can be integrated into any multi-modal intermediate fusion method, the hyper-parameters may be different with each other. We search the appropriate hyper-parameters with two steps: 1) scale search: ranging each hyper-parameter in $\{0.5, 0.05, 0.005, 0.0005, 0.00005\}$, 2) fine-grained search: ranging each hyper-parameter in the scale that achieves the best performance in step 1). For example, if $\lambda_1 = 0.05$ achieves the best performance in step 1), we then range λ_1 from 0.01 to 0.09 in step 2).

Table 6: Hyper-parameters (λ_1, λ_2 and λ_3) settings in our experiments.

METHOD	DATASET	λ_1	λ_2	λ_3	METHOD	DATASET	λ_1	λ_2	λ_3
SELF-MM	CMU-MOSI	0.08	0.05	0.005	MMBT	MVSA-S	0.005	0.5	0.05
	CMU-MOSEI	0.003	0.009	0.08		FOOD101	0.5	0.005	0.5
MMIM	CMU-MOSI	0.08	0.07	0.9	CLMLF	MVSA-S	0.02	0.0004	0.02
	CMU-MOSEI	0.6	0.04	0.0003		HFM	0.5	0.5	0.005
MCL-MCF	CMU-MOSI	0.01	0.3	0.09	MVCN	MVSA-S	0.05	0.05	0.005
	CMU-MOSEI	0.007	0.0003	0.8		HFM	0.5	0.0005	0.1
ATCAF	CMU-MOSI	0.09	0.0005	0.3	URMF	MVSA-S	0.5	0.07	0.09
	CMU-MOSEI	0.02	0.0004	0.007		FOOD101	0.05	0.0005	0.05
MMML	CMU-MOSI	0.005	0.05	0.05	-	-	-	-	-
	CMU-MOSEI	0.05	0.0005	0.05		-	-	-	-

Additionally, we report the sensitive analysis of several methods: Self-MM, MMIM, MCL-MCF and AtCAF on CMU-MOSI dataset. We fix other hyper-parameters to the value in Table 6 when a specific hyper-parameter are ranging in the corresponding scale of magnitude.

Figure 6 shows the variation tendencies of FCD's prediction performance with the changing of values for the hyper-parameters in Equation (17), *i.e.*, (1) λ_1: the weight for $\mathcal{L}_{\text{MI}}$ which supervises the extraction of unique feature without redundant information, (2) λ_2: the weight for $\mathcal{L}_{\text{dis}}$ which keeps the discriminative information between redundant features and further constrains the modality-specific feature extraction, and (3) λ_3: the weight for $\mathcal{L}_{\text{SDA}}$ which controls the quality of modality-invariant information extraction and synergistic feature alignment. From Figure 6, there seems to be no obvious common trend among different methods for each hyper-parameter. This may be caused by the way how FCD cooperates with each method. Generally, *MAE* shows an opposite trend of change compared to other metrics. For most cases, there exists a significant peak (or valley for *MAE*) of each metric, representing the suitable value. When *MAE* reaches a valley, other metrics reach peaks and vice versa. However, for λ_3 of AtCAF, there exists a peak for *MAE* and valleys for other metrics, which is significantly different with other cases. This may be caused by an inappropriate amplitude of change, where a more fine-grained hyper-parameter search is required. Therefore, FCD needs careful hyper-parameter searching to achieve its greatest potential.

F Computational Overhead Analysis

To further analysis the effectiveness of FCD, we give the following computational overhead analysis to discover training time cost brought by FCD. We conduct this experiment employing Self-MM, MMIM, MCL-MCF, AtCAF, and MMML as the base models to calculate the training overhead per epoch (seconds). The results are shown in Table 7.

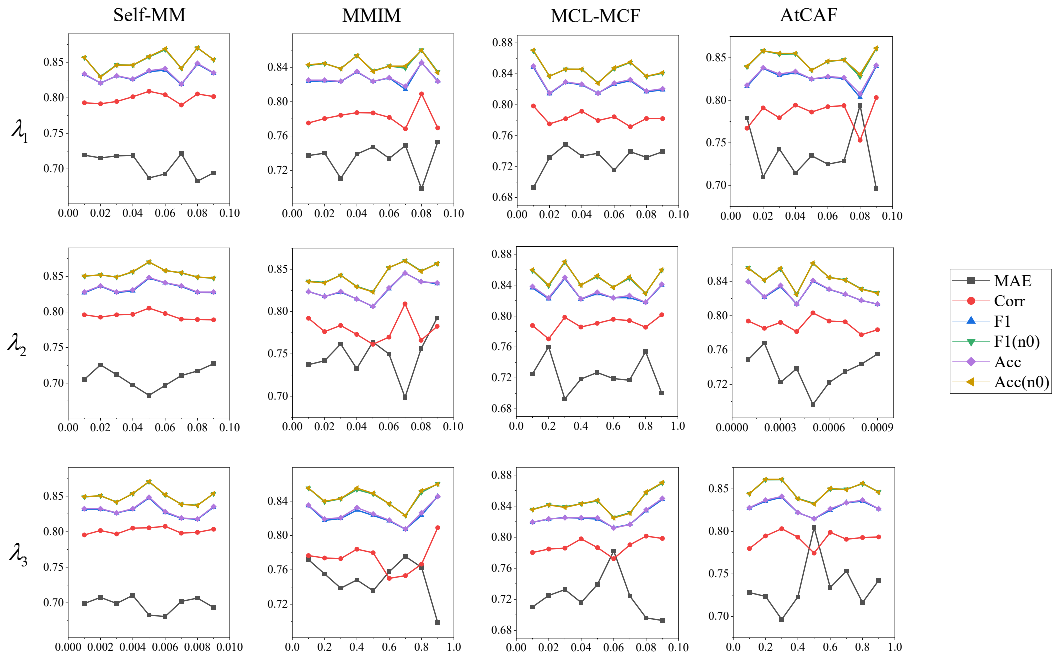


Figure 6: Sensitive analysis of hyper-parameters λ_1, λ_2 and λ_3 with Self-MM, MMIM, MCL-MCF and AtCAF. The horizontal coordinate axis denotes the hyper-parameter rangings and the vertical coordinate axis denotes the evaluation metrics. "n0" in F1(n0) and Acc(n0) stands for the exclusion of 0 when these two metrics are calculated.

Table 7: The computational time overhead per epoch (seconds) of Self-MM, MMIM, MCL-MCF, AtCAF, and MMML.

Method	Dataset	Base (w/o FCD)	Ours (w/ FCD)
Self-MM	CMU-MOSI	3.92±0.40	5.12±0.47
	CMU-MOSEI	37.14±3.72	51.13±4.37
MMIM	CMU-MOSI	18.73±1.30	24.98±1.90
	CMU-MOSEI	97.74±4.61	131.57±8.09
MCF-MCL	CMU-MOSI	21.19±1.43	27.45±2.39
	CMU-MOSEI	242.30±7.87	327.76±10.84
AtCAF	CMU-MOSI	17.26±1.69	21.95±1.70
	CMU-MOSEI	200.61±9.15	250.68±9.97
MMML	CMU-MOSI	228.52±1.37	239.12±1.85
	CMU-MOSEI	2558.47±107.74	2570.14±90.25

In Table 7, we report the average value and the standard deviation of duration to train one epoch on CMU-MOSI and CMU-MOSEI datasets when FCD is applied (Ours (w/ FCD)) or not (Base (w/o FCD)). From Table 7, we can find that the training time of each epoch is different between different methods. This may be caused by the original structure and computation complexity of each base method. When FCD is applied, the increment is also different. This may be caused by the various hidden dimensions, the default batch sizes and the number of modalities (*e.g.*, MMML only has two modalities).

G Broader Impacts and Future Works

FCD is a plug-and-play module that can be integrated into any existing intermediate multimodal models to handle the unimodal uncertain noise whilst makes full use of the task-related information.

We believe that FCD can bring more attentions to current multimodal representation learning community about handling both of the task-related features (*i.e.* the synergistic and unique features) and the unimodal uncertainty noise. However, the quality and semantic richness of unimodal feature is not fully explored. In the training phase of Self-MM and MMML, we find that the prediction performance of each unimodal is quite different. Although there have been researches, such as [43; 5], that engage on estimating the reliability of unimodal prediction, it still remains to be excavated when unimodal uncertain noise is removed for better intermediate fusion. In this case, how much task-related information can unimodal features provide should be considered. In the future, we will make our effort toward this situation to overcome the issues that the quality and semantic richness of unimodal features are various.