IBMA: Information Bottleneck-Based Multi-modal Alignment

Yancheng Wang[1] **Zeyu Dong**[1] **Dongfang Sun**[1] **Alvin C. Silva**[2] **Teresa Wu**[1] **Yingzhen Yang**[1]
[1] School of Computing and Augmented Intelligence, Arizona State University
{ywan1053, zdong112, dsun30, teresa.wu, yingzhen.yang}@asu.edu
[2] Mayo Clinic Arizona
silva.alvin@mayo.edu

Abstract

Multimodal learning aims to integrate information from heterogeneous data sources to improve representation quality and downstream task performance. A key challenge lies in aligning modality-specific representations while suppressing modality-dependent noise and redundancy. The Information Bottleneck (IB) principle provides a principled framework for learning task-relevant representations. Existing multimodal IB methods primarily apply the IB principle to fused multimodal representation and rely on restrictive distributional assumptions, such as Gaussian latent priors induced by variational autoencoders, which may not hold in practice. In this paper, we propose Information Bottleneck–based Multimodal Alignment (IBMA), a novel multimodal learning framework that enforces the IB principle for both the fused multimodal representation and modality-specific representations. IBMA introduces modality-specific representation alignment that guides each modality-specific encoder to learn informative and task-relevant representations aligned with the complementary modality, thereby enhancing cross-modal semantic consistency. Moreover, we derive a novel, efficient, and distribution-free variational upper bound for the IB loss that avoids unrealistic assumptions on latent feature distributions and is readily optimized using standard stochastic gradient descent. Extensive experiments demonstrate that IBMA achieves superior performance compared to existing multimodal learning methods, validating the effectiveness of modality-specific representation alignment. The code for IBMA is available at https://github.com/Statistical-Deep-Learning/IBMA.

1 Introduction

Multimodal learning integrates heterogeneous data sources such as images and text to learn robust representations for downstream tasks (Sleeman et al., 2023; Li et al., 2024; Yuan et al., 2025). Most existing approaches employ separate modality-specific encoders and fuse them via cross-modal attention (Lu et al., 2023; Cai et al., 2023; Wu et al., 2025) or graph neural networks (Zhang et al., 2023; Liu et al., 2025). However, multimodal learning remains challenging due to modality-specific noise, redundancy, and semantic misalignment (Mai et al., 2023a; Wu et al., 2025). While alignment methods such as contrastive learning (Guo et al., 2024; Zou et al., 2023; Boecking et al., 2022; Zhang et al., 2024; Xie et al., 2025) promote cross-modal consistency, they do not explicitly suppress redundant or noisy information within individual modalities.

To explicitly control the trade-off between informativeness and redundancy in multimodal representations, the Information Bottleneck (IB) principle has been introduced as a theoretically grounded framework for multimodal representation learning (Wu et al., 2025; Mai et al., 2023a). The IB principle aims to learn features that are strongly correlated with class labels while reducing their correlation with the inputs. Let X, Z, and Y denote the input features, learned multimodal features, and ground-truth class labels, respectively. The IB principle is enforced by increasing the mutual information between Z and Y while reducing that between Z and X, which corresponds to reducing the IB loss $I(Z, X) - I(Z, Y)$, where $I(\cdot, \cdot)$ denotes mutual information. Motivated by this principle,

several IB-based multimodal learning methods aim to learn discriminative representations by reducing the IB loss or its variational bound (Wu et al., 2025; Mai et al., 2023a).

Limitation of Existing Methods. MIB (Mai et al., 2023a), MCIB (Wang et al., 2025), and OMIB (Wu et al., 2025) explicitly formulate multimodal information bottleneck objectives for fused representations and derive tractable variational upper bounds to explicitly reduce the IB loss. However, both existing IB-based methods (Mai et al., 2023a; Wang et al., 2025; Wu et al., 2025) rely on an unrealistic distributional assumption by adopting a VAE-style surrogate, in which the conditional distribution $p(\zeta \mid z)$ is enforced to match the marginal $p(\zeta)$, which is further approximated by a fixed Gaussian prior. Herein z denotes the learned task-relevant representation, and ζ is a latent feature introduced for multimodal representation learning, obtained by encoding z through a VAE encoder. In addition, MIB (Mai et al., 2023a), MCIB (Wang et al., 2025), and OMIB (Wu et al., 2025) apply the IB principle solely to the fused multimodal representation, aiming to learn an informative joint feature.

To address these limitations, we propose Information Bottleneck–based Multimodal Alignment (IBMA), which enforces the IB principle to both the fused multimodal representation and the individual modality-specific representations to learn informative and task-relevant representations with guidance from the other modality. For each modality, IBMA introduces an information-theoretic regularization term that reduces mutual information with the raw input while increasing mutual information with the representation learned from the complementary modality. This design explicitly encourages modality-specific encoders to focus on task-relevant and semantically aligned information, thus enhancing modality-wise feature alignment and leading to improved performance. In addition, we derive a novel, efficient, and distribution-free variational upper bound for the IB loss, termed IBB, which avoids restrictive assumptions on latent feature distributions (Mai et al., 2023a; Wu et al., 2025). The resulting objective is separable across training samples and can be optimized efficiently using standard SGD-based training procedures. We demonstrate the effectiveness of IBMA for multimodal learning tasks, where multimodal data from different domains must be aligned under significant appearance variation and noise.

As illustrated by the Grad-CAM visualization of the image encoders for multimodal emotion recognition in Figure 1, although the speech scripts of both speakers show clear indications of the ground-truth emotion, the state-of-the-art IB-based multimodal learning method, OMIB (Wu et al., 2025), fails to attend to the most emotion-relevant regions in the images. In particular, OMIB either misses critical facial cues, such as the angry face in the top figure and the smiling mouth in the bottom figure, or mistakenly focuses on background regions, such as the guitar and the arm of the other person in the bottom figure, potentially hurting the performance of multimodal emotion recognition. In contrast, our proposed IBMA focuses on the important parts of the human body with clear indications of the speaker's emotion, such as the angry face and gesture of the person in the top figure and the smiling mouth of the person in the bottom figure, guided by the clear indication of the emotion from the speech script of the speaker. This is because the modality-specific representation alignment proposed in IBMA, as explained earlier, enables textual guidance for image representations, which is absent in existing IB-based multimodal learning methods such as OMIB (Wu et al., 2025). Ablation study in Table 5 further demonstrates the advantages of the modality-specific representation alignment. In addition, Table 6 demonstrates the advantages of our efficient and distribution-free variational upper bound

Figure 1: Grad-CAM visualization of the image encoder from OMIB and IBMA for two instances from the MELD (Poria et al., 2019) dataset for multimodal emotion recognition (anger in the top and joy in the bottom). The speech scripts from the audio are attached below the input images. The speech scripts in both examples clearly convey emotional cues that match the emotions expressed by the speakers. The utterance "Oh God, get out!" reflects strong anger, while "That's great!" clearly expresses joy.

for the IB loss over existing upper bounds for the IB loss (Dai et al., 2018; Guo et al., 2023; Cheng et al., 2020). Please refer to Figure 3 in Section C.6 of the appendix for more visualization results on CheXpert (Irvin et al., 2019) for multimodal disease classification.

Contributions. The contributions of this paper are presented as follows.

First, we introduce IBMA, a multimodal learning framework that enforces modality-specific representation alignment through the Information Bottleneck principle, explicitly guiding each encoder to learn informative and task-relevant representations with cross-modal supervision. In contrast with existing works, which promote the IB principle solely on the fused representation, IBMA performs modality-specific representation alignment by the modality-specific IB principle to learn informative and task-relevant representations with guidance from the other modality, as well as the reduction of the IB loss on the fused multimodal representation, thus enhancing modality-wise feature alignment with improved performance.

Second, we present a novel and provable variational upper bound for the IB loss, termed IBB, which can be optimized by standard SGD algorithms. The IB loss is then reduced by reducing its upper bound, the IBB. Different from existing upper bounds for the IB loss, such as VIB (Dai et al., 2018; Srivastava et al., 2021), which impose an unrealistic Gaussian assumption on the hidden features of DNNs, and APIB (Guo et al., 2023), which reduces only an approximation to the IB loss, IBMA directly reduces a variational upper bound for the IB loss, IBB, without introducing any distributional assumptions on the hidden features or approximation to the IB loss. Moreover, the proposed IBB is computationally efficient with a computational complexity of $\Theta(nT_0 + nC^2)$, where C is the number of classes, n is the number of training samples, and T_0 denotes the computational complexity of a forward and backward pass of the neural network with respect to each training sample. In contrast, the upper bound for the mutual information used to bound the IB loss in CLUB (Cheng et al., 2020), albeit not requiring distributional assumptions on the hidden features of DNNs, requires a substantially higher computational complexity of $\Theta(n^2T_0)$ since $n \gg C^2$ on multimodal learning datasets detailed in Table 1. A composite loss, which combines the IBB and the regular cross-entropy loss, is used to train DNNs for multimodal learning tasks. Table 5 demonstrates that the IB loss of DNNs trained for multimodal learning can be reduced by optimizing such a composite loss. Table 6 demonstrates that the models using IBB achieve substantially better performance than the models using CLUB, VIB, and APIB. We remark that our variational upper bound, as an independent contribution, can be applied to broader discriminative tasks beyond IB-based multimodal learning. Throughout this paper we use $[N]$ to denote all the natural numbers between 1 and N.

2 Related Works

Related Works on Multimodal Learning. Early works perform multimodal alignment by leveraging contrastive objectives or attention modules to enforce cross-modal consistency or similarity, thereby projecting heterogeneous modalities into a shared representation space for joint learning (Guo et al., 2024; Mai et al., 2023b; Zou et al., 2023; Boecking et al., 2022; Cai et al., 2023; Zhang et al., 2024; Xie et al., 2025). However, these methods do not explicitly suppress redundant or noisy information within individual modalities. To explicitly control the trade-off between informativeness and redundancy in multimodal representation, the IB principle has been widely adopted in multimodal learning to learn task-relevant information that is shared across different modalities, while effectively suppressing modality-specific noise and redundant information (Mai et al., 2023a; Wang et al., 2025; Fang et al., 2024; Wu et al., 2025). For instance, DMIB (Fang et al., 2024) proposes a dynamic multimodal fusion framework inspired by the IB principle that preserves label-relevant information via a KL-based sufficiency loss, but it does not explicitly minimize the IB objective, as it increases mutual information with labels without suppressing modality-specific redundancy or noise. MIB (Mai et al., 2023a), MCIB (Wang et al., 2025), and OMIB (Wu et al., 2025) formulate multimodal IB objectives and derive tractable variational upper bounds to directly reduce the IB loss. Nevertheless, these methods rely on a restrictive VAE-style surrogate that enforces the conditional distribution $p(\zeta \mid z)$ to match the marginal $p(\zeta)$, which is further approximated by a fixed Gaussian prior, where z denotes the task-relevant representation and ζ is a latent variable introduced for multimodal learning.

Related Works on Information Bottleneck. The Information Bottleneck (IB) principle (Tishby et al., 2000) provides a theoretical framework for representation learning by encouraging latent variables to preserve task-relevant information while discarding redundant input details. Deep

VIB (Alemi et al., 2017) is the first work to explicitly formulate the IB principle as a training objective for deep neural networks. Motivated by IB, (Lai et al., 2021) introduces a spatial attention mechanism to explicitly reduce the IB loss of attention-weighted features, while (Zhou et al., 2022) shows that self-attention can be interpreted as iterative optimization of the IB objective. Beyond architectural insights, recent theoretical studies (Amjad & Geiger, 2020; Kawaguchi et al., 2023) link IB loss control to improved generalization. Moreover, recent works (Wang et al., 2021; Kuang et al., 2023) also suggest that reducing IB loss enhances adversarial robustness.

3 METHODS

This section presents the proposed Information Bottleneck-Based Multimodal Alignment (IBMA) framework. We first introduce the motivation and formulation of IBMA by defining modality-specific and multimodal information bottleneck objectives, and derive a novel, distribution-free variational upper bound for the IB loss that is efficient and amenable to SGD optimization in Section 3.1. We then describe the overall IBMA architecture, including modality-specific encoders and a cross-modal attention module, and detail the training procedure of IBMA in Section 3.2.

3.1 IBMA: INFORMATION BOTTLENECK-BASED MULTIMODAL ALIGNMENT

Motivation. To learn multimodal representations that are more correlated with the ground-truth label and less correlated with the modality-specific inputs, existing multimodal learning methods based on the IB principle (Mai et al., 2023a; Wu et al., 2025) propose to reduce the multimodal IB loss on the multimodal representation, $\text{IB} = I(Z, X^{(1)}) + I(Z, X^{(2)}) - I(Z, Y)$, where $X^{(j)}$ denotes the random variable representing the input feature in the j-th modality for $j \in \{1, 2\}$. In the following text, the domain superscript j takes value in $\{1, 2\}$. Z denotes the random variable representing the multimodal representation, which is obtained by fusing modality-specific representations $Z^{(j)}$ of each domain $j \in \{1, 2\}$. Y denotes the random variable representing the ground-truth training class label. However, existing IB-based methods (Mai et al., 2023a; Wang et al., 2025; Wu et al., 2025) neglect the complementary information encoded in representations learned from different modalities, which could be exploited to learn modality-invariant semantic information while suppressing modality-specific redundancy and noise. To address this issue, we propose a novel multimodal alignment method, Information Bottleneck-Based Multimodal Alignment (IBMA).

IBMA aims to reduce the modality-specific IB loss on the modality-specific representations, $\text{IB}^{(j)} = I(Z^{(j)}, X^{(j)}) - I(Z^{(j)}, Z^{(j')})$, in addition to reducing the multimodal IB loss. In the following text, $j' \neq j$ denotes the index of the modality other than j, e.g., $j' = (j \bmod 2) + 1$. Reduction of the modality-specific IB loss ensures that each learned modality-specific representation is more semantically correlated with the modality-specific representation from the other domain and less correlated with the input features. Existing IB-based multimodal learning methods (Mai et al., 2023a; Wu et al., 2025), which do not explicitly reduce the modality-specific IB loss, may retain modality-specific noise or redundant information that is not semantically aligned with the other modality. By reducing the modality-specific IB loss, IBMA explicitly encourages learning task-relevant and modality-invariant representations in each modality, while suppressing modality-specific noise and redundancy. This is achieved by increasing the correlation among learned modality-specific representations across domains while decreasing their correlation with the input features, ultimately leading to more discriminative multimodal representations fused from more discriminative modality-specific representations.

Given the multimodal training data $\{X_i^{(1)}, X_i^{(2)}, Y_i\}_{i=1}^n$ from two different modalities, where $X_i^{(j)}$ stands for the i-th input feature from the j-th modality, and Y_i is the ground-truth class label for $i \in [n]$, we first specify how to compute the modality-specific IB loss based on the training data. Let $\text{IB}^{(j)} = I(Z^{(j)}, X^{(j)}) - I(Z^{(j)}, Z^{(j')})$ denote the modality-specific IB loss on the j-th domain, which treats the representation from the other domain as a cross-modal supervision, guiding $Z^{(j)}$ to retain only the most informative signal of the input from domain j semantically aligned with the other domain j'. $X^{(j)}$ takes values in $\{X_i^{(j)}\}_{i=1}^n$. $Z^{(j)}$ takes values in $\{Z_i^{(j)}\}_{i=1}^n$ with $Z_i^{(j)}$ being the i-th learned representation by the modality-specific encoder on the j-th modality. Y takes values in $\{Y_i\}_{i=1}^n$. We have the class centroids $\{\mathcal{F}_a^{(j)}\}_{a=1}^C$ and $\{\mathcal{C}_y^{(j)}\}_{y=1}^C$ for the learned representations $\{Z_i^{(j)}\}_{i=1}^n$ and the input features $\{X_i^{(j)}\}_{i=1}^n$ for the j-th domain. Each $\mathcal{F}_a^{(j)}$ is the average of the

modality-specific representation in class a, and each $\mathcal{C}_y^{(j)}$ is the average of the input features in class y, where C is the number of classes. Then we define the probability that $Z_i^{(j)}$ belongs to class a as $\Pr\left[Z^{(j)} \in a\right] = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i^{(j)}, a)$ with $\phi(Z_i^{(j)}, a) = \frac{\exp\left(-\left\|Z_i^{(j)} - \mathcal{F}_a^{(j)}\right\|_2^2\right)}{\sum_{a'=1}^{C}\exp\left(-\left\|Z_i^{(j)} - \mathcal{F}_{a'}^{(j)}\right\|_2^2\right)}$. We define the probability that $X_i^{(j)}$ belongs to class y as $\Pr\left[X^{(j)} \in y\right] = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i^{(j)}, y)$. Similarly, we define the joint probabilities $\Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right] = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i^{(j)} \in a, X_i^{(j)} \in y)$ and $\Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i^{(j)} \in a, Z_i^{(j')} \in y)$, where $\phi(Z_i^{(j)} \in a, X_i^{(j)} \in y)$ and $\phi(Z_i^{(j)} \in a, Z_i^{(j')} \in y)$ are defined in (3) and (4) in the appendix. As a result, we can compute the mutual information $I(Z^{(j)}, X^{(j)})$ and $I(Z^{(j)}, Z^{(j')})$ by

$$I(Z^{(j)}, X^{(j)}) = \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right] \log \frac{\Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right]}{\Pr\left[Z^{(j)} \in a\right]\Pr\left[X^{(j)} \in y\right]},$$

$$I(Z^{(j)}, Z^{(j')}) = \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{\Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right]}{\Pr\left[Z^{(j)} \in a\right]\Pr\left[Z^{(j')} \in y\right]},$$

and then compute the IB loss $\mathrm{IB}^{(j)}$. Given a variational distribution $Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)$ for $y \in [C]$ and $a \in [C]$, the following theorem gives a variational upper bound, $\mathrm{IBB}^{(j)}$, for the IB loss $\mathrm{IB}^{(j)}$ on the j-th domain.

Theorem 3.1. *Let* $\Pr\left[X^{(j)} \in y\right] = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = y\}}/n := p_y$ *for every* $y \in [C]$. *Then the IB loss on the* j*-th modality satisfies*

$$\mathrm{IB}^{(j)} \leq \mathrm{IBB}^{(j)}, \tag{1}$$

where the variational upper bound is given by

$$\mathrm{IBB}^{(j)} := \frac{1}{n}\sum_{i=1}^{n}\left(U_i^{(j)} - V_i^{(j)}\right).$$

Here

$$U_i^{(j)} := \sum_{a=1}^{C}\sum_{y=1}^{C} \phi_j(i,a,y) \log\left(\frac{\phi_j(i,a,y)}{p_y \phi(Z_i^{(j)}, a)}\right), V_i^{(j)} := \sum_{a=1}^{C}\sum_{y=1}^{C} \phi_{jj'}(i,a,y) \log Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y),$$

where $\phi_j(i,a,y) = \phi(Z_i^{(j)} \in a, X_i^{(j)} \in y)$, $\phi_{jj'}(i,a,y) = \phi(Z_i^{(j)} \in a, Z_i^{(j')} \in y)$. $Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)$ is the variational conditional probability that $Z^{(j)}$ belongs to class a given $Z^{(j')}$ belongs to class y, which is computed efficiently by Algorithm 3 in the supplementary. The proof of Theorem 3.1 follows by applying Lemma A.1 and Lemma A.2 in Section A of the supplementary. We remark that $\mathrm{IBB}^{(j)}$ is ready to be optimized by standard SGD algorithms because it is separable and expressed as the summation of losses on individual training points. Algorithm 1 describes the training process of an IBMA model. The following functions are needed for minibatch-based training with SGD, with the subscript b indicating the corresponding loss on the b-th batch $\mathcal{B}_b$:

$$\mathrm{IBB}_b^{(j)} = \frac{1}{|\mathcal{B}_b|}\sum_{i \in \mathcal{B}_b}\left(U_i^{(j)} - V_i^{(j)}\right).$$

In addition to reducing the modality-specific IB loss $\mathrm{IB}^{(j)}$ by reducing $\mathrm{IBB}^{(j)}$, IBMA also reduces the multimodal IB loss on the multimodal representation Z fused from the modality-specific representations. Let $\mathrm{IB}^{(\text{fuse})} = I(Z, X^{(1)}) + I(Z, X^{(2)}) - I(Z, Y)$ be the multimodal IB loss on the multimodal representation Z, where Y denotes the class label taking value in $\{Y_i\}_{i=1}^{n}$. $\mathrm{IBB}^{(\text{fuse})}$ denotes the variational upper bound for $\mathrm{IB}^{(\text{fuse})}$ which is derived in a similar manner as in Theorem 3.1 and to be detailed in Section A.2 of the appendix. Here Z takes values in $\{Z_i\}_{i=1}^{n}$ and Z_i denotes the fused multimodal representation for the i-th input. We compute class centroids $\{\mathcal{F}_a\}_{a=1}^{C}$ on the fused

multimodal representation, where each $\mathcal{F}_a$ is the average of the fused multimodal representation in class a. We also define the probability that Z_i belongs to class a as $\Pr[Z \in a] = \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i, a)$ similar to the computation of $\Pr\left[Z^{(j)} \in a\right]$. The training loss of IBMA on the b-th batch is then computed by

$$\mathcal{L}_b = \mathrm{CE}_b + \eta \mathrm{IBB}_b, \tag{2}$$

where $\mathrm{IBB}_b = \mathrm{IBB}_b^{(1)} + \mathrm{IBB}_b^{(2)} + \mathrm{IBB}_b^{(\mathrm{fuse})}$ and $\mathrm{CE}_b = \mathrm{CE}_b^{(1)} + \mathrm{CE}_b^{(2)} + \mathrm{CE}_b^{(\mathrm{fuse})}$. $\mathrm{CE}_b^{(j)} = \frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b} H(Z_i^{(j)}, Y_i)$ is the cross-entropy loss on the b-th batch $\mathcal{B}_b$ in the j-th domain. $\mathrm{CE}_b^{(\mathrm{fuse})} = \frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b} H(Z_i, Y_i)$ is the cross-entropy loss on the b-th batch $\mathcal{B}_b$ for the fused multimodal representation. $H(\cdot,\cdot)$ is the cross-entropy function. η is the balance factor for IBB. The values of η on different datasets will be decided by performing cross-validation.

Efficient and Distribution-Free Variational Upper Bound. Because our variational upper bounds are distribution-free, we demonstrate the computational efficiency of $\mathrm{IBB}^{(j)}$ over the existing distribution-free upper bound for the IB loss in CLUB (Cheng et al., 2020), and a similar conclusion holds for $\mathrm{IBB}^{(\mathrm{fuse})}$. Let T_0 denote the computational complexity of a forward and backward pass of the neural network with respect to each training sample. The overall computational complexity for calculating the proposed modality-specific IBB regularization term $\mathrm{IBB}^{(j)}$ is $\Theta(nT_0 + nC^2)$. In contrast, computing the upper bound for the mutual information required for calculating the upper bound for the IB loss proposed in CLUB (Cheng et al., 2020) requires a substantially higher computational complexity at least $\Theta(n^2T_0)$ since $n \gg C^2$ on multimodal learning datasets detailed in Table 1, which covers the benchmarks in the extensive literature (Wu et al., 2025; Mai et al., 2023a) where $C \leq 7$. We note that $\Theta(n^2T_0)$ corresponds exclusively to the upper bound for the mutual information $I(Z, X^{(j)})$, while CLUB additionally requires the computation of the lower bound for the mutual information $I(Z, Y)$. Details on the complexity analysis of CLUB and our IBB are presented in Section D of the supplementary. In contrast to existing IB upper bounds that rely on the unrealistic Gaussian distributional assumption of hidden features (Dai et al., 2018; Srivastava et al., 2021) or on the approximation to the IB loss (Guo et al., 2023), our bound for the IB loss, the IBB, does not require any distributional assumptions on the hidden features or approximation to the IB loss. Table 6 in Section 4.6 demonstrates that IBB achieves substantially better performance than CLUB and other competing methods that rely on the unrealistic Gaussian distribution assumption of hidden features (Dai et al., 2018) or on the approximation to the IB loss (Guo et al., 2023), and requires much less training time than the distribution-free baseline method, CLUB (Cheng et al., 2020).

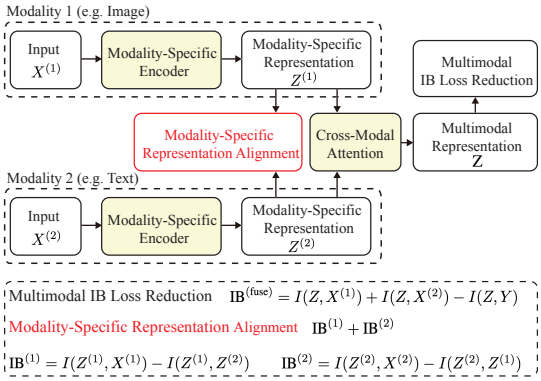

Figure 2: IBMA employs modality-specific encoders to learn representations from each input modality, such as image and text. A cross-modal attention module integrates the modality-specific representations to form a fused multimodal representation. In addition to reducing the multimodal IB loss on the fused multimodal representation, IBMA enforces modality-specific representation alignment by leveraging representations from other modalities as cross-modal supervision.

3.2 IBMA Architecture and Training Procedure

The proposed IBMA framework comprises two main components, namely modality-specific encoders that generate modality-specific representations and a cross-modal attention module that fuses modality-specific representations into a multimodal representation. An overview of the architecture is illustrated in Figure 2.

Model Architecture of IBMA. Given the input $X_i^{(j)}$ from the j-th domain for the i-th input, IBMA employs a modality-specific encoder $f^{(j)}(\cdot)$ to extract modality-specific representation $Z_i^{(j)} = f^{(j)}(X_i^{(j)})$. Each modality-specific encoder is implemented using a modality-specific backbone,

such as ResNet (He et al., 2016) for image data and BERT (Devlin et al., 2019) for text data. The modality-specific representations $Z_i^{(1)}$ and $Z_i^{(2)}$ for the i-th input are fused using a cross-modal attention module widely used in the multimodal learning literature (Wu et al., 2025; Wang et al., 2025), which produces a joint multimodal representation $Z_i = F(Z_i^{(1)}, Z_i^{(2)})$, where $F(\cdot,\cdot)$ denotes the cross-modal attention module. The fused representation Z_i for the i-th input is used for the downstream tasks. IBMA naturally extends to tasks with more than two modalities by applying a modality-specific encoder to each modality and performing cross-modal attention–based fusion jointly across multiple modalities following (Mai et al., 2023a; Wu et al., 2025), while applying modality-specific representation alignment to each pair of modalities.

Algorithm 1 Training Algorithm of the IBMA Model

Input: Multimodal training data $\left\{X_i^{(1)}, X_i^{(2)}, Y_i\right\}_{i=1}^{n}$, mini-batches $\{\mathcal{B}_b\}_{b=1}^{B}$, total training epochs t_{train}, learning rate α.
Output: Trained network parameters $\mathcal{W}$.

```
Initialize the network parameters W randomly.
for j = 1 to 2 do
    Compute initial class centroids {C_y^(j)}_{y=1}^C.
    j' = (j mod 2) + 1
    Initialize the variational distribution Q^(j)(Z^(j) ∈ a | Z^(j') ∈ y) = 1/C for all a, y ∈ [C].
end for
Initialize the fused variational distribution Q(Z ∈ a | Y = y) = 1/C for all a, y ∈ [C].
Initialize fused class centroids {F_a}_{a=1}^C for Z as zeros or using one forward pass.
for t = 1 to t_train do
    for b = 1 to B do
        Update W ← W − α∇_W L_b using mini-batch B_b, where L_b is defined in Equation (2).
    end for
    for j = 1 to 2 do
        j' = (j mod 2) + 1
        Update Q^(j)(Z^(j) ∈ a | Z^(j') ∈ c) using Algorithm 3.
        Update class centroids {F_a^(j)}_{a=1}^C.
    end for
    Update Q(Z ∈ a | Y = y) using Algorithm 4.
    Update class centroids {F_a}_{a=1}^C.
end for
return Trained network parameters W.
```

Training Procedure of IBMA. IBMA is trained in an end-to-end manner by jointly optimizing all model components, including the modality-specific encoders and the cross-modal fusion module. The network parameters are updated by optimizing the joint training loss $\mathcal{L}_b$ in Equation (2) for the training data in the b-th batch. Algorithm 1 describes the training of the IBMA model. Such a joint optimization strategy ensures that IBMA simultaneously achieves modality-wise feature alignment and effective multimodal fusion under the guidance of the IB principle.

4 EXPERIMENT

In this section, we evaluate the proposed IBMA across multiple multimodal learning tasks. Section 4.1 first describes the datasets and experimental settings throughout the experiments. Section 4.2 evaluates IBMA for multimodal emotion recognition. Section 4.3 evaluates the performance of IBMA for multimodal sentiment analysis. Section 4.4 evaluates the performance of IBMA on multimodal disease classification on two large-scale datasets, MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019). Section 4.5 presents an ablation study analyzing the impact of modality-specific representation alignment in IBMA. Section 4.6 compares the proposed IBB variational upper bound for the IB loss with existing IB upper-bound formulations in terms of both effectiveness and training efficiency. Additional experiment results are presented in Section C of the appendix. Section C.1 evaluates the effectiveness of IBMA for multimodal anomalous tissue detection. We analyze the sensitivity of IBMA to the choice of η in Section C.2. Section C.3 reports the training time comparisons between IBMA and the competing multimodal learning baselines.

Section C.4 evaluates the statistical significance of the observed performance improvements of IBMA in Table 2-6. We study the classification performance using the modality-specific encoder alone in IBMA in Section C.5. The Grad-CAM visualization analysis on CheXpert for multimodal disease classification is presented in Section C.6.

4.1 Implementation Details

We evaluate IBMA on various multimodal benchmarks, including emotion recognition datasets, CREMA-D (Cao et al., 2014), MELD (Poria et al., 2019), and IEMOCAP (Busso et al., 2008), sentiment analysis datasets, CMU-MOSI (Zadeh et al., 2016) and MELD (Poria et al., 2019), pathological tissue datasets, 10x-hNB-A–H and 10x-hBC-A–D (Xu et al., 2024), and disease classification datasets, MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019). Five common thorax diseases from MIMIC-CXR and CheXpert are used following the literature on thorax disease classification (Xiao et al., 2023). Detailed dataset descriptions are provided in Section B.1 of the appendix, and Table 1 summarizes the training and test splits for all datasets used in our experiments. Detailed experimental settings and descriptions of the compared methods are provided in Section B.2 and B.3 of the appendix.

Table 1: Summary of training and test data sizes for each dataset.

Dataset	# Classes	Training size	Test size
CREMA-D (Cao et al., 2014)	6	6,698	744
CMU-MOSI (Zadeh et al., 2016)	7	1,281	685
10x-hNB & 10x-hBC (Xu et al., 2024)	2	18,183	1,123
IEMOCAP (Busso et al., 2008)	5	5,990	1,541
MELD (Poria et al., 2019)	7	9,989	2,610
MIMIC-CXR (Johnson et al., 2019)	5	227,835	1,000
CheXpert (Irvin et al., 2019)	5	223,414	234

4.2 Multimodal Emotion Recognition

We evaluate IBMA on multimodal emotion recognition, where emotional states are jointly predicted from heterogeneous modalities, following standard protocols in the multimodal learning literature (Wu et al., 2025; Mai et al., 2023a; Cui et al., 2024). For CREMA-D, we adopt the audio–visual setting in (Wu et al., 2025), where raw speech waveforms are converted into log-Mel spectrograms and encoded using a ResNet-18 backbone to obtain fixed-dimensional acoustic embeddings, while facial video frames are processed by a shared-weight ResNet-18 encoder and temporally averaged to produce utterance-level visual representations. For MELD and IEMOCAP, we adopt a tri-modal setting with audio, visual, and textual modalities following (Mai et al., 2023a), where modality-specific representation alignment is applied to each pair of modalities. The textual modality is encoded using a BERT-based text encoder (Liu et al., 2019). The resulting modality-specific representations are fused via a cross-modal attention module to produce a unified multimodal representation for emotion prediction. It is observed in Table 2 that IBMA consistently outperforms existing multimodal learning methods across all three benchmarks. In particular, IBMA outperforms the strongest baseline by 1.8% on CREMA-D, demonstrating the effectiveness of modality-specific representation alignment.

Table 2: Accuracy comparison for emotion recognition on CREMA-D, MELD, and IEMOCAP. The IBMA results are averaged over 10 runs with different random initializations and reported as mean $\pm$ standard deviation, while the statistical significance of its improvements over the best baseline is presented in Table 10 in Section C.4 of the appendix.

Method	CREMA-D	MELD	IEMOCAP
Concat (Perez et al., 2018)	53.2	56.1	67.4
BiGated (Kiela et al., 2018)	58.4	59.8	70.2
MISA (Hazarika et al., 2020)	57.7	60.5	71.0
Deep IB (Wang et al., 2019)	54.1	57.3	68.6
MMIB (Zhang et al., 2022)	56.7	61.4	72.3
MMRLIB (Cui et al., 2024)	57.3	62.0	73.1
MIB (Mai et al., 2023a)	61.4	63.6	74.1
MCIB (Wang et al., 2025)	63.2	64.0	<u>74.3</u>
CLFA (Zhang et al., 2024)	63.0	63.5	74.0
DCLF (Xie et al., 2025)	63.4	63.2	74.2
OMIB (Wu et al., 2025)	<u>63.6</u>	<u>64.3</u>	74.2
IBMA (Ours)	**65.4**$\pm$0.4	**65.7**$\pm$0.3	**75.7**$\pm$0.3

4.3 Multimodal Sentiment Analysis

We evaluate IBMA on multimodal sentiment analysis, where sentiment polarity is jointly inferred from heterogeneous modalities, following standard protocols in the multimodal learning literature (Wu et al., 2025; Mai et al., 2023a; Cui et al., 2024). For both CMU-MOSI and MELD, we adopt a tri-modal setting following (Wu et al., 2025), incorporating audio, visual, and textual modalities. Raw speech waveforms are transformed into log-Mel spectrograms and encoded with a ResNet-18 backbone, while facial video frames are processed by a shared-weight ResNet-18 encoder with temporal average pooling. Textual inputs are encoded using a BERT-based text encoder (Liu et al., 2019). The resulting modality-specific representations are fused through a cross-modal attention module to obtain a unified multimodal representation for sentiment prediction. We report seven-class accuracy (Acc-7) on CMU-MOSI to evaluate fine-grained sentiment polarity and intensity, and additionally report binary accuracy (Acc-2) and F1-score by grouping sentiments into positive and negative classes. For MELD, we report Acc-2 and F1-score only, as it provides binary sentiment annotations. As shown in Table 3, IBMA consistently outperforms existing multimodal learning methods on both benchmarks. In particular, IBMA surpasses the strongest baseline by 1.4% in F1-score and 1.5% in binary accuracy on MELD, demonstrating the effectiveness of modality-specific representation alignment.

Table 3: Performance comparison for sentiment analysis on CMU-MOSI and MELD. The IBMA results are averaged over 10 runs with different random initializations and reported as mean $\pm$ standard deviation, while the statistical significance of its improvements over the best baseline is presented in Table 11 in Section C.4 of the appendix.

Method	CMU-MOSI			MELD	
	Acc-7	Acc-2	F1	Acc-2	F1
Concat (Perez et al., 2018)	41.5	81.1	82.0	74.2	73.8
BiGated (Kiela et al., 2018)	41.8	82.1	83.2	75.1	74.6
MISA (Hazarika et al., 2020)	42.3	83.4	83.6	76.3	75.8
Deep IB (Wang et al., 2019)	45.3	83.2	83.3	76.0	75.5
MMIB (Zhang et al., 2022)	46.3	85.0	85.0	78.4	78.0
MMRLIB (Cui et al., 2024)	45.7	84.3	84.4	77.8	77.3
DMIB (Fang et al., 2024)	40.4	83.2	83.3	75.6	75.2
MIB (Mai et al., 2023a)	48.6	85.3	85.3	79.1	78.7
MCIB (Wang et al., 2025)	48.8	86.7	87.2	80.2	79.7
CLFA (Zhang et al., 2024)	47.8	86.1	86.6	80.2	79.7
DCLF (Xie et al., 2025)	48.5	86.5	87.1	80.1	79.5
OMIB (Wu et al., 2025)	48.6	86.9	87.1	80.5	80.1
IBMA (Ours)	**50.1**$\pm$0.2	**87.9**$\pm$0.3	**88.3**$\pm$0.2	**82.0**$\pm$0.4	**81.5**$\pm$0.3

4.4 Multimodal Disease Classification

We evaluate IBMA for multimodal thoracic disease classification on the large-scale medical vision-language benchmarks MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019), where the goal is to predict the presence of multiple thoracic pathologies from the chest X-ray and the clinical reports. Chest X-ray images are encoded with a ViT-B (Xiao et al., 2023) encoder and radiology reports with a BERT-based text encoder (Huang et al., 2019) pre-trained for medical predictions. The same modality-specific encoders and fusion architecture are used across all baseline methods. For both MIMIC-CXR and CheXpert, we report mean AUROC (mAUC), defined as the arithmetic mean of per-disease AUROC scores across all diseases. It is observed in Table 4 that IBMA significantly outperforms all competing multimodal learning methods. For instance, IBMA outperforms OMIB by 1.4% in mAUC on CheXpert, demonstrating the advantages of IBMA.

4.5 Ablation Study on the Modality-Specific Representation Alignment

To study the effect of modality-specific representation alignment in IBMA, we construct an ablation model, termed IBMA w/o Modality-Specific Representation Alignment, which removes all modality-specific IB loss while retaining the same modality-specific encoders, cross-modal fusion module, and multimodal IB loss on the fused multimodal representation. We compare IBMA with the ablation model and OMIB (Wu et al., 2025) in terms of modality-specific IB loss, multimodal IB loss, and the mAUC for multimodal disease classification. It is observed in Table 5 that IBMA renders significantly lower modality-specific IB loss than OMIB and the ablation model without modality-specific representation alignment, leading to a lower multimodal IB loss and better classification accuracy. For instance, the modality-specific IB losses of IBMA are 0.022 and 0.013 lower than those of the ablation model without modality-specific representation alignment, leading to a 1% improvement in mAUC on CheXpert.

Table 4: mAUC comparison for disease classification on MIMIC-CXR and CheXpert. The IBMA results are averaged over 10 runs with different random initializations and reported as mean $\pm$ standard deviation, while the statistical significance of its improvements over the best baseline is presented in Table 12 in Section C.4 of the appendix.

Method	MIMIC-CXR	CheXpert
Concat (Perez et al., 2018)	67.9	85.0
BiGated (Kiela et al., 2018)	68.4	87.5
MISA (Hazarika et al., 2020)	68.2	86.4
Deep IB (Wang et al., 2019)	67.0	86.2
MMIB (Zhang et al., 2022)	69.6	87.7
MMRLIB (Cui et al., 2024)	69.4	87.5
MIB (Mai et al., 2023a)	70.2	88.1
MCIB (Wang et al., 2025)	70.6	88.4
CLFA (Zhang et al., 2024)	70.5	88.6
DCLF (Xie et al., 2025)	71.0	88.9
OMIB (Wu et al., 2025)	71.0	89.3
IBMA (Ours)	**72.2**$\pm$0.3	**90.7**$\pm$0.2

Table 5: Ablation study on the modality-specific representation alignment for multimodal disease classification. Lower (more negative) values of IB loss indicate stronger IB alignment and better adherence to the IB principle. The IBMA results are averaged over 10 runs with different random initializations and reported as mean $\pm$ standard deviation, while the statistical significance of its improvements over the ablation model is presented in Table 15 in Section C.4 of the appendix.

Dataset	Method	Visual IB Loss $\downarrow$	Textual IB Loss $\downarrow$	Multimodal IB Loss $\downarrow$	mAUC (%) $\uparrow$
MIMIC-CXR	OMIB (Wu et al., 2025)	-0.021	-0.024	-0.028	71.0
	IBMA w/o Modality-Specific Representation Alignment	-0.028	-0.031	-0.035	71.4
	IBMA (Ours)	**-0.048**	**-0.045**	**-0.058**	**72.2**$\pm$0.3
CheXpert	OMIB (Wu et al., 2025)	-0.020	-0.023	-0.028	89.3
	IBMA w/o Modality-Specific Representation Alignment	-0.027	-0.030	-0.036	89.7
	IBMA (Ours)	**-0.049**	**-0.043**	**-0.055**	**90.7**$\pm$0.2

4.6 Comparison between IBB and Existing Upper Bounds for the IB Loss

We compare the proposed variational upper bound for the IB loss, IBB, with existing works deriving the upper bound for the IB loss (Cheng et al., 2020; Dai et al., 2018; Srivastava et al., 2021), including CLUB (Cheng et al., 2020) and VIB (Dai et al., 2018; Srivastava et al., 2021). We also compare IBB with APIB (Guo et al., 2023), which relies on IB approximation instead of directly reducing the IB loss. The comparison is conducted by replacing IBB with VIB, APIB, and CLUB in IBMA for emotion recognition on CREMA-D, with training time reported on a single NVIDIA A100 GPU. It is observed in Table 6 that the unrealistic Gaussian distribution assumption on the hidden features imposed by VIB (Dai et al., 2018) and the IB approximation in APIB (Guo et al., 2023) lead to degraded performance compared with IBB. For example, the IBB-based model outperforms the VIB-based model by 1.5% in accuracy on CREMA-D, with only a marginal 5.3% increase in training time. Moreover, the IBB-based model outperforms the CLUB-based model by 1.3% in accuracy while using only 52.6% of its training time.

Table 6: Comparison of different methods for reducing the IB loss on CREMA-D for emotion recognition. The IBB results are averaged over 10 runs with different random initializations and reported as mean $\pm$ standard deviation, while the statistical significance of its improvements over the best baseline is presented in Table 14 in Section C.4 of the appendix.

Methods	Training Time (Minutes/Epoch)	Accuracy
VIB (Dai et al., 2018)	1.9	63.9
APIB (Guo et al., 2023)	1.9	63.8
CLUB (Cheng et al., 2020)	3.8	64.1
IBB (Ours)	2.0	**65.4**$\pm$0.4

5 Conclusion

In this paper, we propose Information Bottleneck–based Multimodal Alignment (IBMA), a novel multimodal learning framework that enforces the IB principle at both the fused multimodal and modality-specific representation levels. By introducing modality-specific representation alignment, IBMA explicitly guides each modality-specific encoder to learn task-relevant and semantically aligned representations with cross-modal supervision, while effectively suppressing modality-dependent noise and redundancy. In addition, we derive a novel, efficient, and distribution-free variational upper bound for the IB loss, IBB. Extensive experiments and ablation studies across diverse multimodal benchmarks demonstrate that IBMA significantly outperforms existing multimodal learning methods.

ACKNOWLEDGMENTS

This work was supported by National Institute of Health 1OT2OD037955-01. This work was also supported by the 2023 Mayo Clinic and Arizona State University Alliance for Health Care Collaborative Research Seed Grant Program under Award No. AWD00038846.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Rana Ali Amjad and Bernhard C. Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9):2225–2239, 2020.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Wetscherek, Tristan Naumann, Aditya V. Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pp. 1–21. Springer, 2022.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008.

Gan Cai, Yu Zhu, Yue Wu, Xiaoben Jiang, Jiongyao Ye, and Dawei Yang. A multimodal transformer to fuse images and metadata for skin disease classification. *Vis. Comput.*, 39(7):2781–2793, 2023.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.*, 5(4):377–390, 2014.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1779–1788. PMLR, 2020.

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:1274–1285, 2024.

Bin Dai, Chen Zhu, Baining Guo, and David P. Wipf. Compressing neural networks using the variational information bottleneck. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1143–1152. PMLR, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, pp. 1–17, 2025.

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. Dynamic multimodal information bottleneck for multimodality classification. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 7681–7691. IEEE, 2024.

Song Guo, Lei Zhang, Xiawu Zheng, Yan Wang, Yuchao Li, Fei Chao, Chenglin Wu, Shengchuan Zhang, and Rongrong Ji. Automatic network pruning via hilbert-schmidt independence criterion lasso under information bottleneck principle. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 17412–17423. IEEE, 2023.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (eds.), *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pp. 1122–1131. ACM, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 590–597. AAAI Press, 2019.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16049–16096. PMLR, 2023.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomás Mikolov. Efficient large-scale multimodal classification. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5198–5204. AAAI Press, 2018.

Huafeng Kuang, Hong Liu, Yongjian Wu, Shin'ichi Satoh, and Rongrong Ji. Improving adversarial robustness via information bottleneck distillation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Qiuxia Lai, Yu Li, Ailing Zeng, Minhao Liu, Hanqiu Sun, and Qiang Xu. Information bottleneck approach to spatial attention learning. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 779–785. ijcai.org, 2021.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Comput. Biol. Medicine*, 177:108635, 2024.

Liang Liu, Shuaiyong Li, and Yongqiang Zhu. Audio-visual semantic graph network for audio-visual event localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 23957–23966. Computer Vision Foundation / IEEE, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Comput. Sci.*, 9: e1400, 2023.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Trans. Multim.*, 25:4121–4134, 2023a.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.*, 14(3):2276–2289, 2023b.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3942–3951. AAAI Press, 2018.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 527–536. Association for Computational Linguistics, 2019.

Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4390–4399. PMLR, 2018.

William C. Sleeman, Rishabh Kapoor, and Preetam Ghosh. Multimodal classification: Current landscape, taxonomy and future directions. *ACM Comput. Surv.*, 55(7):150:1–150:31, 2023.

Ayush Srivastava, Oshin Dutta, Jigyasa Gupta, Sumeet Agarwal, and Prathosh AP. A variational information bottleneck based method to compress sequential networks for human action recognition. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pp. 2744–2753. IEEE, 2021.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.),

Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008, 2017.

Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. Deep multi-view information bottleneck. In Tanya Y. Berger-Wolf and Nitesh V. Chawla (eds.), *Proceedings of the 2019 SIAM International Conference on Data Mining, SDM 2019, Calgary, Alberta, Canada, May 2-4, 2019*, pp. 37–45. SIAM, 2019.

Yihua Wang, Qi Jia, Cong Xu, Feiyu Chen, Yuhan Liu, Haotian Zhang, Liang Jin, Lu Liu, and Zhichun Wang. Conditional information bottleneck for multimodal fusion: Overcoming shortcut learning in sarcasm detection. *CoRR*, abs/2508.10644, 2025.

Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, and Jennifer G. Dy. Revisiting hilbert-schmidt information bottleneck for adversarial robustness. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 586–597, 2021.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3588–3600, 2023.

Yunhe Xie, Chengjie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. A dual contrastive learning framework for enhanced multimodal conversational emotion recognition. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 4055–4065. Association for Computational Linguistics, 2025.

Kaichen Xu, Yan Lu, Suyang Hou, Kainan Liu, Yihang Du, Mengqian Huang, Hao Feng, Hao Wu, and Xiaobo Sun. Detecting anomalous anatomic regions in spatial transcriptomics with stands. *Nature Communications*, 15(1):8223, 2024.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Comput. Surv.*, 57(7):167:1–167:34, 2025.

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

Ming Zhang, Ke Chang, and Yunfang Wu. Multi-modal semantic understanding with contrastive cross-modal feature alignment. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 11934–11943. ELRA and ICCL, 2024.

Tonghui Zhang, Haiying Zhang, Shuke Xiang, and Tong Wu. Information bottleneck based representation learning for multimodal sentiment analysis. In *CCEAI 2022: The 6th International Conference on Control Engineering and Artificial Intelligence, Virtual Event, Japan, March 11 - 13, 2022*, pp. 7–11. ACM, 2022.

Yanteng Zhang, Xiaohai He, Yi Hao Chan, Qizhi Teng, and Jagath C. Rajapakse. Multi-modal graph neural network for early diagnosis of alzheimer's disease from smri and PET scans. *Comput. Biol. Medicine*, 164:107328, 2023.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and José M. Álvarez. Understanding the robustness in vision transformers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27378–27394. PMLR, 2022.

Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. Unis-mmc: Multimodal classification via unimodality-supervised multimodal contrastive learning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 659–672. Association for Computational Linguistics, 2023.

A Proof of Theorem 3.1

We define

$$\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right) = \frac{\exp\left(-\left\|Z_i^{(j)} - \mathcal{F}_a^{(j)}\right\|_2^2 - \left\|X_i^{(j)} - \mathcal{C}_y^{(j)}\right\|_2^2\right)}{\sum_{a'=1}^{C}\sum_{y'=1}^{C}\exp\left(-\left\|Z_i^{(j)} - \mathcal{F}_{a'}^{(j)}\right\|_2^2 - \left\|X_i^{(j)} - \mathcal{C}_{y'}^{(j)}\right\|_2^2\right)}, a \in [C], y \in [C], \tag{3}$$

$$\phi\left(Z_i^{(j)} \in a, Z_i^{(j')} \in y\right) = \frac{\exp\left(-\left\|Z_i^{(j)} - \mathcal{F}_a^{(j)}\right\|_2^2 - \left\|Z_i^{(j')} - \mathcal{F}_y^{(j')}\right\|_2^2\right)}{\sum_{a'=1}^{C}\sum_{y'=1}^{C}\exp\left(-\left\|Z_i^{(j)} - \mathcal{F}_{a'}^{(j)}\right\|_2^2 - \left\|Z_i^{(j')} - \mathcal{F}_{y'}^{(j')}\right\|_2^2\right)}, a \in [C], y \in [C]. \tag{4}$$

Lemma A.1. *Let* $\Pr\left[X^{(j)} \in y\right] = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=y\}}/n := p_y$ *for every* $y \in [C]$*, then*

$$I(Z^{(j)}, X^{(j)}) \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right)\log\left(\frac{\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right)}{p_y\phi(Z_i^{(j)}, a)}\right). \tag{5}$$

Proof. Then the joint probability $\Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right]$ is specified by

$$\Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right] = \frac{1}{n}\sum_{i=1}^{n}\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right). \tag{6}$$

It then follows from (6) and the log sum inequality that

$$\begin{aligned}
I(Z^{(j)}, X^{(j)}) &= \sum_{a=1}^{C}\sum_{y=1}^{C}\Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right]\log\frac{\Pr\left[Z^{(j)} \in a, X^{(j)} \in y\right]}{\Pr\left[Z^{(j)} \in a\right]\Pr\left[X^{(j)} \in y\right]} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right)\log\left(\phi\left(Z^{(j)} \in a, X^{(j)} \in y\right)\right) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right)\log\left(\phi(Z_i^{(j)}, a)p_y\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right)\log\left(\frac{\phi\left(Z_i^{(j)} \in a, X_i^{(j)} \in y\right)}{p_y\phi(Z_i^{(j)}, a)}\right).
\end{aligned} \tag{7}$$

□

Lemma A.2. *Recall that* $j' = (j \bmod 2) + 1$ *is the modality other than* j*, we have*

$$I(Z^{(j)}, Z^{(j')}) \geq \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i^{(j)} \in a, Z_i^{(j')} \in y\right)\log Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y). \tag{8}$$

Proof. Let $Q^{(j)}(Z^{(j)}|Z^{(j')})$ be a variational distribution. We have

$$\begin{aligned}
I(Z^{(j)}, Z^{(j')}) &= \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{\Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right]}{\Pr\left[Z^{(j)} \in a\right] \Pr\left[Z^{(j')} \in y\right]} \\
&= \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{\Pr\left[Z^{(j)} \in a | Z^{(j')} \in y\right] Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)}{\Pr\left[Z^{(j)} \in a\right] Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)} \\
&\geq \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{\Pr\left[Z^{(j)} \in a | Z^{(j')} \in y\right]}{Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)} \\
&\quad + \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)}{\Pr\left[Z^{(j)} \in a\right]} \\
&= \mathrm{KL}\left(P(Z^{(j)}|Z^{(j')}) \middle\| Q^{(j)}(Z^{(j)}|Z^{(j')})\right) \\
&+ \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)}{\Pr\left[Z^{(j)} \in a\right]} \\
&\geq \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log \frac{Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)}{\Pr\left[Z^{(j)} \in a\right]} \\
&= \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y) + H\left(P(Z^{(j)})\right) \\
&\geq \sum_{a=1}^{C}\sum_{y=1}^{C} \Pr\left[Z^{(j)} \in a, Z^{(j')} \in y\right] \log Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y) \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C} \phi\left(Z_i^{(j)} \in a, Z_i^{(j')} \in y\right) \log Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y). \qquad (9)
\end{aligned}$$

□

Proof of Theorem 3.1. Equation (1) in Theorem 3.1 of the main paper follows from IB $= I(Z^{(j)}, X^{(j)}) - I(Z^{(j)}, Z^{(j')})$, the upper bound for $I(Z^{(j)}, X^{(j)})$ in Lemma A.1 and the lower bound for $I(Z^{(j)}, Z^{(j')})$ in Lemma A.2. □

A.1 Computation of $Q^{(j)}(Z^{(j)}|Z^{(j')})$

The variational distribution $Q^{(j)}(Z^{(j)}|Z^{(j')})$ is computed by

$$Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y) = \Pr\left[Z^{(j)} \in a | Y = y\right] = \frac{\sum_{i=1}^{n} \phi(Z_i^{(j)}, a) \mathbb{1}_{\{Y_i = y\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Y_i = y\}}}, \quad a, y \in [C]. \quad (10)$$

Our Algorithm 3 computes $Q^{(j)}(Z^{(j)}|Z^{(j')})$ efficiently with a time complexity of $\Theta(nC + nT_0)$, where C is the number of classes, n is the number of training samples, and T_0 denotes the computational complexity of a forward and backward pass of the neural network with respect to each training sample.

We remark that the variational distribution $Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y) = \Pr\left[Z^{(j)} \in a | Y = y\right]$ is modeled through the conditional probability on the class label $Y = y$. We first emphasize that our upper bound IBB$^{(j)}$ in Theorem 3.1 for the IB loss IB$^{(j)}$ holds for an arbitrary distribution

$Q^{(j)}(Z^{(j)} \in a | Z^{(j')} \in y)$. Our particular choice for $Q^{(j)}$ is not only effective in empirical performance as evidenced in Section 4, but also efficient which eliminates of the need of a time consuming process of training a VAE to model the conditional distribution $P(Z^{(j)}|Z^{(j')})$, such as CLUB (Cheng et al., 2020).

A.2 Computation of $\text{IBB}^{(\text{fuse})}$

Repeating the argument in the proof of Theorem 3.1, the upper bound for the IB loss on the fused multimodal representation is $\text{IB}^{(\text{fuse})} \leq \text{IBB}^{(\text{fuse})}$, where

$$\begin{aligned}
\text{IBB}^{(\text{fuse})} = & \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i \in a, X_i^{(j)} \in y\right)\log\left(\frac{\phi\left(Z_i \in a, X_i^{(1)} \in y\right)}{p_y\phi(Z_i, a)}\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi\left(Z_i \in a, X_i^{(2)} \in y\right)\log\left(\frac{\phi\left(Z_i \in a, X_i^{(2)} \in y\right)}{p_y\phi(Z_i, a)}\right) \\
& - \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{C}\sum_{y=1}^{C}\phi(Z_i, a)\mathbb{1}_{\{Y_i=y\}}\log Q(Z \in a | Y = y), \qquad (11)
\end{aligned}$$

where the variational distribution $Q(Z \in a|Y = y)$ is computed by

$$Q(Z \in a|Y = y) = \Pr\left[Z \in a|Y = y\right] = \frac{\sum_{i=1}^{n}\phi(Z_i, a)\mathbb{1}_{\{Y_i=y\}}}{\sum_{i=1}^{n}\mathbb{1}_{\{Y_i=y\}}}, \quad a, y \in [C]. \qquad (12)$$

Our Algorithm 4 computes $Q(Z|Y)$ efficiently with a time complexity of $\Theta(nC + nT_0)$, where C is the number of classes, n is the number of training samples, and T_0 denotes the computational complexity of a forward and backward pass of the neural network with respect to each training sample.

B Additional Experiment Settings

B.1 Datasets

For speech-based emotion recognition, we adopt CREMA-D (Cao et al., 2014), which consists of acted audio-visual recordings where professional actors convey six categorical emotions through synchronized facial expressions and speech signals. In addition, we include IEMOCAP (Busso et al., 2008), a widely used conversational emotion corpus featuring dyadic interactions with fine-grained emotional annotations, allowing us to further assess robustness under spontaneous and context-dependent emotional expressions. For multimodal sentiment analysis (MSA), experiments are conducted on CMU-MOSI (Zadeh et al., 2016), which integrates language, acoustic, and visual modalities and provides real-valued sentiment intensity labels ranging from -3 (strongly negative) to +3 (strongly positive). We further extend sentiment evaluation to MELD (Poria et al., 2019), a large-scale multimodal dialogue dataset derived from television series, where utterance-level emotion and sentiment labels are annotated within multi-speaker conversational contexts. To study multimodal learning in biomedical settings, we consider eight pathological tissue datasets constructed from healthy human breast tissues (10x-hNB-A–H) and human breast cancer tissues (10x-hBC-A-D) (Xu et al., 2024). Each dataset jointly contains gene expression profiles and corresponding histology images. Following prior work (Wu et al., 2025; Mai et al., 2023a), IBMA is trained exclusively on healthy tissue datasets and subsequently transferred to cancer tissue datasets for pathological tissue detection, enabling evaluation under a cross-domain generalization setting. In addition, we involve large-scale multimodal learning datasets, MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019), which are large-scale chest radiography datasets consisting of paired medical images and clinical text reports, and are widely used benchmarks for multimodal representation learning and medical image understanding.

B.2 Experiment Settings

IBMA employs modality-specific encoders and a cross-modal attention–based fusion architecture for multimodal learning across tasks, where each encoder produces a 512 dimensional embedding as specified in the corresponding experimental sections. The cross-modal fusion module is implemented using a single transformer block (Vaswani et al., 2017) with 8 attention heads and a position-wise MLP, together with residual connections and layer normalization, after which the fused features are projected to a 512 dimensional multimodal representation following (Wu et al., 2025). All models are trained for 500 epochs using the Adam optimizer, with the learning rate selected from $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ and the weight decay from $\{1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}\}$. The balancing factor η for the IBB regularization term is selected via five-fold cross-validation from $\{0.1, 0.2, \ldots, 0.9\}$ on the training set of each evaluation benchmark.

B.3 Compared Methods

We compare IBMA with a broad range of multimodal fusion and information-theoretic baselines, including early fusion method, Concat (Perez et al., 2018) and BiGated (Kiela et al., 2018), representation disentanglement method, MISA (Hazarika et al., 2020), and information bottleneck–based methods, Deep IB (Wang et al., 2019), MMIB (Zhang et al., 2022), MMRLIB (Cui et al., 2024), MIB (Mai et al., 2023a), MCIB (Wang et al., 2025), and OMIB (Wu et al., 2025), and contrastive learning based alignment methods, DCLF (Xie et al., 2025) and CLFA (Zhang et al., 2024). The best results among the three variants of MIB (Mai et al., 2023a), including E-MIB, L-MIB, and C-MIB, are reported.

C Additional Experiment Results

C.1 Anomalous Tissue Detection

We further evaluate IBMA on multimodal anomalous tissue detection, where the objective is to identify anomalous tissue regions from eight human breast cancer datasets (10x-hBC-A–D) by jointly modeling gene expression and histology modalities. Due to the limited availability of region-level anomaly annotations, we adopt the Support Vector Data Description (SVDD) framework (Ruff et al., 2018) for anomaly detection. Following OMIB (Wu et al., 2025), we employ modality-specific encoders, where gene expression profiles are encoded using a multi-layer perceptron (MLP) and histology regions are encoded using a ResNet-18 encoder. The resulting modality-specific representations are fused using the cross-modal attention module, leading to the fused multimodal representation. The same modality-specific encoders are used across all baseline methods to ensure a fair comparison with IBMA. The model is trained exclusively on eight healthy tissue datasets (10x-hNB-A–H) to learn a compact hypersphere that encloses multimodal representations of normal tissue regions. During inference, the trained model is applied to each breast cancer dataset, and the distance between each multimodal representation and the hypersphere center is used as an anomaly score to identify anomalous regions. All competing methods follow the same SVDD-based training and evaluation protocol and use identical modality-specific encoders for fair comparison. Detection performance is evaluated using ROC-AUC and F1-score, computed based on the predicted anomaly scores and available region-level annotations. It is observed in Table 7 that IBMA significantly outperforms the competing multimodal learning methods on anomalous tissue detection. For instance, IBMA outperforms the strongest baseline, OMIB, by 2% in AUC averaged on all four test datasets.

C.2 Sensitivity Analysis

We analyze the sensitivity of IBMA to the balancing factor η in the joint training loss $\mathcal{L}_b$ from Equation (2) on CREMA-D, where η controls the strength of the IBB regularization relative to the cross-entropy loss. We conduct the sensitivity study by varying η over the range $\{0.1, 0.2, \ldots, 0.9\}$ while keeping all other training settings fixed. As shown in Table 8, IBMA maintains stable performance across different values of η, with an accuracy variation of at most 0.4% on CREMA-D for multimodal emotion recognition.

Table 7: Performance comparison of multimodal learning methods for anomalous tissue detection on the 10x-hBC-{A–D} datasets. The IBMA results are averaged over 10 runs with different random initializations and reported as mean $\pm$ standard deviation, while the statistical significance of its improvements over the best baseline is presented in Table 13 in Section C.4 of the appendix.

Method	10x-hBC-A		10x-hBC-B		10x-hBC-C		10x-hBC-D		Mean	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Concat (Perez et al., 2018)	0.537	0.884	0.866	0.654	0.638	0.750	0.555	0.509	0.649	0.699
BiGated (Kiela et al., 2018)	0.489	0.821	0.518	0.352	0.563	0.727	0.540	0.494	0.528	0.599
MISA (Hazarika et al., 2020)	0.498	0.873	0.499	0.213	0.586	0.754	0.495	0.450	0.520	0.573
Deep IB (Wang et al., 2019)	0.522	0.878	0.379	0.102	0.433	0.693	0.484	0.443	0.455	0.529
MMIB (Zhang et al., 2022)	0.623	0.894	0.818	0.559	0.765	0.822	0.501	0.465	0.677	0.685
MMRLIB (Cui et al., 2024)	0.626	0.897	0.817	0.583	0.662	0.783	0.604	0.524	0.677	0.697
DMIB (Fang et al., 2024)	0.423	0.865	0.849	0.607	0.743	0.827	0.642	0.540	0.664	0.710
MIB (Mai et al., 2023a)	0.598	0.891	0.770	0.483	0.659	0.786	0.652	0.564	0.602	0.681
MCIB (Wang et al., 2025)	0.684	0.898	0.852	0.628	0.716	0.811	0.621	0.548	0.718	0.721
CLFA (Zhang et al., 2024)	0.663	0.892	0.834	0.612	0.704	0.803	0.610	0.541	0.703	0.712
DCLF (Xie et al., 2025)	0.701	0.900	0.872	0.641	0.728	0.819	0.632	0.552	0.733	0.728
OMIB (Wu et al., 2025)	0.728	0.904	0.903	0.663	0.743	0.820	0.640	0.561	0.754	0.737
IBMA (Ours)	**0.744**$\pm$0.003	**0.915**$\pm$0.002	**0.916**$\pm$0.002	**0.685**$\pm$0.005	**0.767**$\pm$0.004	**0.840**$\pm$0.003	**0.668**$\pm$0.004	**0.577**$\pm$0.004	**0.774**	**0.754**

Table 8: Sensitivity of IBMA to the balancing coefficient η on CREMA-D for multimodal emotion recognition.

η	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Accuracy (%)	65.1	65.0	65.2	65.3	**65.4**	65.3	65.1	65.0	65.0

C.3 Training Time Comparison

We further compare the training efficiency of IBMA with existing multimodal learning methods on the CREMA-D dataset. All methods are trained on a single NVIDIA A100 GPU. The training time results on CREMA-D are summarized in Table 9, where the total training time for both IBMA and the baseline methods is reported. It is observed in Table 9 that IBMA only marginally increases the training time by 5.3% compared to the strongest baseline OMIB, while significantly outperforming OMIB by 1.8% in emotion recognition accuracy on CREMA-D, demonstrating the effectiveness and efficiency of the proposed IBMA. It is worthwhile to mention that the class centroids in the input feature space can be efficiently pre-computed prior to training using FAISS (Douze et al., 2025), which supports fast, memory-efficient, and highly parallelized centroid and distance computations on GPUs, even for high-dimensional features. In particular, computing the class centroids and the distances between the input features of all samples in the training set of CREMA-D and their corresponding class centroids requires only 16.7 seconds, which happens one time before the training. Computing the class centroids and the distances between the learned features of all samples in the training set of CREMA-D and their corresponding class centroids requires only 2.1 seconds in each epoch.

Table 9: Performance and training time comparison of multimodal learning methods on CREMA-D.

Method	CREMA-D	
	Accuracy (%)	Training Time (Minutes/Epoch)
Concat (Perez et al., 2018)	53.2	1.5
BiGated (Kiela et al., 2018)	58.4	1.7
MISA (Hazarika et al., 2020)	57.7	1.8
Deep IB (Wang et al., 2019)	54.1	1.8
MMIB (Zhang et al., 2022)	56.7	1.9
MMRLIB (Cui et al., 2024)	57.3	2.2
MIB (Mai et al., 2023a)	61.4	1.8
MCIB (Wang et al., 2025)	63.2	1.9
CLFA (Zhang et al., 2024)	63.0	2.0
DCLF (Xie et al., 2025)	63.4	2.0
OMIB (Wu et al., 2025)	63.6	1.9
IBMA (Ours)	**65.4**$\pm$0.4	2.0

C.4 Improvement Significance Analysis

To verify that the performance improvements achieved by IBMA over existing multimodal learning methods are statistically significant and not caused by random variation, we conduct paired t-tests between IBMA and the best-performing baseline method on each dataset. For all the multimodal learning tasks, IBMA and the corresponding best baseline are trained and evaluated 10 times with different random seeds for network initialization. The mean and standard deviation of the evaluation metrics, together with the resulting p-values of the paired t-tests, are reported in Table 10 for emotion

recognition, Table 11 for sentiment analysis, Table 12 for multimodal disease classification, and Table 13 for anomalous tissue detection. As shown in Tables 10–13, the largest p-value across all benchmarks and tasks is 3.6×10^{-5} on MELD for emotion recognition, which is below the standard significance threshold of 0.05. These results indicate that the performance improvements achieved by IBMA over the corresponding best baseline methods are statistically significant with $p \ll 0.05$, and are unlikely to be caused by random variation.

Table 10: Statistical significance analysis for multimodal emotion recognition. Paired t-tests are conducted between IBMA and the best baseline on each dataset over 10 runs.

Dataset	Best Baseline (Accuracy)	IBMA (Accuracy)	p-value
CREMA-D	63.6 ± 0.3	65.4 ± 0.4	1.2×10^{-6}
MELD	64.3 ± 0.4	65.7 ± 0.3	3.6×10^{-5}
IEMOCAP	74.3 ± 0.3	75.7 ± 0.3	4.5×10^{-6}

Table 11: Statistical significance analysis for multimodal sentiment analysis. Paired t-tests are conducted between IBMA and the best baseline on each dataset over 10 runs.

Dataset	Best Baseline (Acc-2)	IBMA (Acc-2)	p-value
CMU-MOSI	86.9 ± 0.3	87.9 ± 0.3	2.2×10^{-7}
MELD	80.5 ± 0.4	82.0 ± 0.4	1.9×10^{-6}

Table 12: Statistical significance analysis for multimodal disease classification. Paired t-tests are conducted between IBMA and the best baseline on each dataset over 10 runs.

Dataset	Best Baseline (mAUC)	IBMA (mAUC)	p-value
MIMIC-CXR	71.0 ± 0.2	72.2 ± 0.3	3.6×10^{-6}
CheXpert	89.3 ± 0.2	90.7 ± 0.2	1.1×10^{-6}

Table 13: Statistical significance analysis for multimodal anomalous tissue detection. Paired t-tests are conducted between IBMA and the best baseline over 10 runs.

Dataset	Best Baseline (AUC)	IBMA (AUC)	p-value
10x-hBC-A	0.728 ± 0.004	0.744 ± 0.003	2.5×10^{-7}
10x-hBC-B	0.903 ± 0.002	0.916 ± 0.002	1.6×10^{-6}
10x-hBC-C	0.743 ± 0.003	0.767 ± 0.004	5.7×10^{-7}
10x-hBC-D	0.652 ± 0.004	0.668 ± 0.004	4.2×10^{-7}

To further verify that the performance improvements brought by IBB over existing IB loss upper bounds are statistically significant rather than due to random variation, we additionally conduct paired t-tests between the IBB-based model and each competing variant. All models are trained and evaluated 10 times on CREMA-D with different random seeds under identical experimental settings. As shown in Table 14, the performance gains achieved by IBB over VIB, APIB, and CLUB are all statistically significant, with p-values well below the standard significance threshold of 0.05. These results further confirm that the superiority of IBB is consistent and robust, and cannot be attributed to stochastic training effects.

Table 14: Statistical significance analysis for comparing IBB with the best-performing IB loss upper bound on CREMA-D. Paired t-tests are conducted over 10 runs with different random seeds.

Dataset	Best Baseline (Accuracy)	IBB (Accuracy)	p-value
CREMA-D	64.1 ± 0.3	65.4 ± 0.4	2.8×10^{-7}

To further verify that the improvements brought by modality-specific representation alignment are statistically significant rather than caused by random variation, we additionally conduct paired t-tests between IBMA and the ablation model without modality-specific representation alignment. Both models are trained and evaluated 10 times with different random seeds under identical experimental settings. The resulting p-values for multimodal disease classification are reported in Table 15.

C.5 Analysis on the Modality-Specific Encoder

To explicitly assess the quality of modality-specific representations learned by IBMA, we evaluate each modality-specific encoder independently on the CheXpert dataset. Following the standard evaluation protocols in existing works (Mai et al., 2023a), we attach a linear classifier to the frozen

Table 15: Statistical significance analysis comparing IBMA with the ablation model without modality-specific representation alignment. Paired t-tests are conducted over 10 runs with different random seeds.

Dataset	IBMA w/o Modality-Specific Representation Alignment (mAUC)	IBMA (mAUC)	p-value
MIMIC-CXR	71.4 ± 0.3	72.2 ± 0.3	3.1×10^{-6}
CheXpert	89.7 ± 0.2	90.7 ± 0.2	1.4×10^{-6}

image encoder and text encoder, respectively, and train the classifier to predict thoracic disease labels using only a single modality at a time. We compare IBMA with OMIB (Wu et al., 2025) and an ablation variant, IBMA without modality-specific representation alignment, which removes all modality-specific IB losses in IBMA. It is observed in Table 16 that IBMA consistently improves the performance of both the image encoder and the text encoder compared to the competing methods. In particular, IBMA achieves significantly better mAUC than OMIB for both modalities, while removing modality-specific representation alignment leads to degraded unimodal performance. For instance, IBMA outperforms OMIB by 1.2% and 1.5% in mAUC for disease classification on the image domain and text domain, eventually leading to an improvement of 1.4% in mAUC for multimodal disease classification. The consistent improvements across both modalities demonstrate that both modality-specific encoders benefit from the cross-domain supervision enforced by the modality-specific representation alignment proposed in IBMA.

Table 16: Evaluation of modality-specific encoders on CheXpert. Each encoder is evaluated independently using a linear classifier on top of frozen representations. Higher mAUC indicates better modality-specific representation quality.

Method	mAUC		
	Image	Text	Multimodal
OMIB (Wu et al., 2025)	88.0	76.0	89.3
IBMA w/o Modality-Specific Representation Alignment	88.4	76.5	89.7
IBMA (Ours)	**89.2**	**77.5**	**90.7**

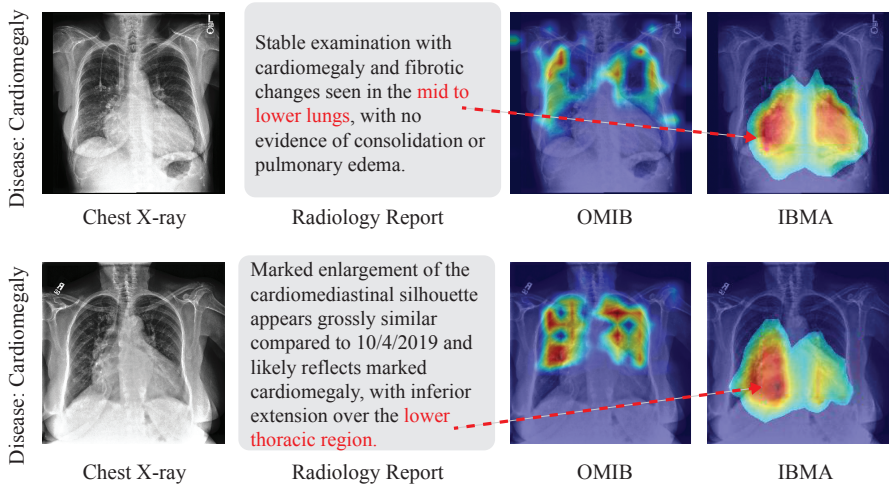

Figure 3: Grad-CAM visualization of the image encoder from OMIB and IBMA for two instances from the CheXpert (Irvin et al., 2019) dataset for multimodal disease classification. The two examples are in the class of cardiomegaly, which is characterized by abnormal enlargement of the heart, commonly reflected on chest X-rays as an increased cardiothoracic ratio observed in the lower middle of the two lungs. The radiology reports in both examples explicitly indicate the presence of cardiomegaly, using phrases such as "mid to lower lungs" and "lower thoracic region", which correspond to the typical anatomical location of cardiomegaly in chest radiographs. The red arrows illustrate that the image encoder of IBMA successfully focuses on the most informative regions of cardiomegaly in the mid to lower or lower lungs, as indicated in the radiology reports.

Algorithm 2 MI Minimization with vCLUB (Algorithm 1 in CLUB (Cheng et al., 2020))

for each training iteration **do**
  Sample $\{(X_i, Z_i)\}_{i=1}^n$ from $p_\sigma(X, Z)$
  Compute log-likelihood $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \log q_\theta(Z_i|X_i)$
  Update $q_\theta(Z|X)$ by increasing $\mathcal{L}(\theta)$
  for $i = 1$ to n **do**
    $L_i = \log q_\theta(Z_i|X_i) - \frac{1}{n}\sum_{j=1}^n \log q_\theta(Z_j|X_i)$
  end for
  Update $p_\sigma(X, Z)$ by reducing $\widehat{I}_{\text{vCLUB}} = \frac{1}{n}\sum_{i=1}^n L_i$
end for

Algorithm 3 Efficient Computation of $Q^{(j)}(Z^{(j)} \in a \mid Z^{(j')} \in y)$

Input: The precomputed soft assignments $\phi(Z_i^{(j)}, a)$ for $i \in [n]$ and $a \in [C]$, the class labels $\{Y_i\}_{i=1}^n$, and the number of classes C. Here $\left\{Z_i^{(j)}\right\}_{i=1}^n$ are computed at the t-th epoch in Algorithm 1 of the main paper.

Output: Conditional distribution matrix $Q \in \mathbb{R}^{C\times C}$

Initialize $Q \leftarrow \mathbf{0}^{C\times C}$ and count vector $M \leftarrow \mathbf{0}^C$.
for $i = 1$ to n **do**
  for $a = 1$ to C **do**
    $Q[a, Y_i] \leftarrow Q[a, Y_i] + \phi(Z_i^{(j)}, a)$
  end for
  $M[Y_i] \leftarrow M[Y_i] + 1$
end for
for $y = 1$ to C **do**
  for $a = 1$ to C **do**
    $Q[a, y] \leftarrow Q[a, y]/M[y]$
  end for
end for
Return Q

C.6 Grad-CAM Visualization for Multimodal Disease Classification

As illustrated by the Grad-CAM visualization of the image encoders for multimodal disease classification in Figure 3, although the radiology reports in both examples provide clear location-specific indications of cardiomegaly, the state-of-the-art IB-based multimodal learning method, OMIB (Wu et al., 2025), fails to consistently attend to the most disease-relevant regions in the mid to lower lungs and instead focuses on the upper lung regions. In contrast, our proposed IBMA accurately concentrates on the clinically meaningful regions associated with cardiomegaly as described in the radiology reports, which are the mid to lower regions in the lung, guided by the explicit textual cues. This is because the modality-specific representation alignment proposed in IBMA enables effective cross-modal textual guidance for image representations, which is absent in existing IB-based multimodal learning methods such as OMIB (Wu et al., 2025).

D Computational Complexity Analysis

Given samples $\{(X_i, Z_i)\}_{i=1}^n$ drawn from the joint distribution $p_\sigma(X, Z) = p_\sigma(Z|X)p(X)$, the goal of CLUB is to optimize the parameters of the predictive neural network $p_\sigma(Z|X)$ such that the resulting joint model $p_\sigma(X, Z)$ induces minimal mutual information between X and Z. Here, $p_\sigma(Z|X)$ serves as the main predictive model, such as an encoding model or a classification model, while $p(X)$ denotes the empirical data distribution. In addition, $q_\theta(Z|X)$ is a variational conditional distribution, also implemented as a neural network, used to approximate $p(Z|X)$ and to provide a differentiable estimator of mutual information. The optimization alternates between updating $q_\theta(Z|X)$ by increasing the log-likelihood $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \log q_\theta(Z_i|X_i)$ to improve the approximation accuracy, and updating $p_\sigma(X, Z)$ by reducing the upper bound for the mutual information, $I(X, Z)$, $\widehat{I}_{\text{vCLUB}} = \frac{1}{n}\sum_{i=1}^n L_i$, where $L_i = \log q_\theta(Z_i|X_i) - \frac{1}{n}\sum_{j=1}^n \log q_\theta(Z_j|X_i)$. Through this alternating process, the algorithm jointly learns an accurate variational estimator q_θ and a predictive model p_σ. The complete training procedure is summarized in Algorithm 2.

Algorithm 4 Efficient Computation of $Q(Z \in a \mid Y = y)$

Input: The precomputed soft assignments $\phi(Z_i, a)$ for $i \in [n]$ and $a \in [C]$, the class labels $\{Y_i\}_{i=1}^n$, and the number of classes C. Here $\{Z_i\}_{i=1}^n$ are computed at the t-th epoch in Algorithm 1 of the main paper.
Output: Conditional distribution matrix $Q \in \mathbb{R}^{C \times C}$

```
Initialize Q ← 0^{C×C} and count vector M ← 0^C.
for i = 1 to n do
    for a = 1 to C do
        Q[a, Y_i] ← Q[a, Y_i] + φ(Z_i, a)
    end for
    M[Y_i] ← M[Y_i] + 1
end for
for y = 1 to C do
    for a = 1 to C do
        Q[a, y] ← Q[a, y]/M[y]
    end for
end for
Return Q
```

Suppose the computation of $q_\theta(Z|X)$ costs T_q and that of $p_\sigma(Z|X)$ costs T_p. Computing $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \log q_\theta(Z_i|X_i)$ requires $\Theta(nT_q)$ time. The computation of vCLUB $\widehat{I}_{\text{vCLUB}} = \frac{1}{n}\sum_{i=1}^n \left[\log q_\theta(Z_i|X_i) - \frac{1}{n}\sum_{j=1}^n \log q_\theta(Z_j|X_i)\right]$ requires computing $[\log q_\theta(Z_j|X_i)]$ for $i, j \in [n]$, which requires $\Theta(n^2 T_q)$ time. Updating p_σ for reducing $\widehat{I}_{\text{vCLUB}}$ over the same n inputs adds nT_p time. Hence, one training epoch requires $\Theta(n^2 T_q + nT_q + nT_p) = \Theta(n^2 T_q + nT_p)$.

Computational Complexity of IBB$^{(j)}$. Herein we analyze the computational complexity for computing the variational upper bound IBB$^{(j)}$ for the modality-specific IB loss. Let T_0 denote the complexity of a forward and backward computation of the model predicting $Z_i^{(j)}$ for one sample. For each epoch, the the computation of $\{Z_i^{(j)}\}_{i=1}^n$ using the modality-specific encoder takes $\Theta(nT_0)$ time. Given $\{Z_i^{(j)}\}_{i=1}^n$, we first pre-compute $\phi(Z_i^{(j)}, a)$ for $i \in [n]$ and $a \in [C]$, which takes $\Theta(nC)$ time.

To compute IBB$^{(j)} = \frac{1}{n}\sum_{i=1}^n \left(U_i^{(j)} - V_i^{(j)}\right)$, we separately compute $\frac{1}{n}\sum_{i=1}^n U_i^{(j)}$, which is the upper bound for $I(Z^{(j)}, X^{(j)})$, and $\frac{1}{n}\sum_{i=1}^n V_i^{(j)}$, which is the lower bound for $I(Z^{(j)}, Z^{(j')})$.

Let $A_{ia}^{(j)} := \exp(-\|Z_i^{(j)} - \mathcal{F}_a^{(j)}\|_2^2)$ and $B_{iy}^{(j)} := \exp(-\|X_i^{(j)} - \mathcal{C}_y^{(j)}\|_2^2)$ for all $a \in [C]$ and $y \in [C]$. Let $D_i^{(j)} := \sum_{a'=1}^C \sum_{y'=1}^C A_{ia'}^{(j)} B_{iy'}^{(j)} = \widehat{A}_i^{(j)} \widehat{B}_i^{(j)}$, where $\widehat{A}_i^{(j)} := \sum_{a'=1}^C A_{ia'}^{(j)}$ and $\widehat{B}_i^{(j)} := \sum_{y'=1}^C B_{iy'}^{(j)}$. For each $i \in [n]$, we first pre-compute all $A_{ia}^{(j)}$ and $B_{iy}^{(j)}$ for $a \in [C]$ and $y \in [C]$, which takes $\Theta(C)$ time. For each $i \in [n]$, we then pre-compute $\widehat{A}_i^{(j)}$, $\widehat{B}_i^{(j)}$, and $D_i^{(j)}$, which takes $\Theta(C)$ time. Then $U_i^{(j)}$ can be computed as

$$
\begin{aligned}
U_i^{(j)} &= \sum_{a=1}^C \sum_{y=1}^C \phi(Z_i^{(j)} \in a, X_i^{(j)} \in y) \log\left(\frac{\phi(Z_i^{(j)} \in a, X_i^{(j)} \in y)}{p_y \phi(Z_i^{(j)}, a)}\right) \\
&= \frac{1}{D_i^{(j)}} \sum_{a=1}^C \sum_{y=1}^C A_{ia}^{(j)} B_{iy}^{(j)} \log\left(\frac{A_{ia}^{(j)} B_{iy}^{(j)} / D_i^{(j)}}{p_y \phi(Z_i^{(j)}, a)}\right) \\
&= -\log D_i^{(j)} - \frac{\widehat{A}_i^{(j)}}{D_i^{(j)}} \sum_{y=1}^C B_{iy}^{(j)} \log p_y + \frac{1}{D_i^{(j)}} \left(\sum_{a=1}^C A_{ia}^{(j)} \log A_{ia}^{(j)}\right) \widehat{B}_i^{(j)} \\
&\quad + \frac{\widehat{A}_i^{(j)}}{D_i^{(j)}} \left(\sum_{y=1}^C B_{iy}^{(j)} \log B_{iy}^{(j)}\right) - \frac{\widehat{B}_i^{(j)}}{D_i^{(j)}} \left(\sum_{a=1}^C A_{ia}^{(j)} \log \phi(Z_i^{(j)}, a)\right).
\end{aligned}
$$

For each $i \in [n]$, the computation of $\frac{\widehat{A}_i^{(j)}}{D_i^{(j)}} \sum_{y=1}^{C} B_{iy}^{(j)} \log p_y$, $\frac{1}{D_i^{(j)}} \left(\sum_{a=1}^{C} A_{ia}^{(j)} \log A_{ia}^{(j)} \right) \widehat{B}_i^{(j)}$, $\frac{\widehat{A}_i^{(j)}}{D_i^{(j)}} \left(\sum_{y=1}^{C} B_{iy}^{(j)} \log B_{iy}^{(j)} \right)$, and $\frac{\widehat{B}_i^{(j)}}{D_i^{(j)}} \left(\sum_{a=1}^{C} A_{ia}^{(j)} \log \phi(Z_i^{(j)}, a) \right)$ each takes $\Theta(C)$ time. As a result, the computation of $U_i^{(j)}$ for each $i \in [n]$ takes $\Theta(C)$ time. Therefore, the computation of $\frac{1}{n} \sum_{i=1}^{n} U_i^{(j)}$ takes $\Theta(nC)$ time.

Given $\phi(Z_i^{(j)}, a)$ for all $i \in [n]$ and $a \in [C]$, the conditional distribution matrix $Q^{(j)}(Z^{(j)} \in a \mid Z^{(j')} \in y) \in \mathbb{R}^{C \times C}$ can be efficiently computed following Algorithm 3. We denote by $Q^{(j)}[a, y] = Q^{(j)}(Z^{(j)} \in a \mid Z^{(j')} \in y)$ the (a, y)-th entry of $Q^{(j)}$, representing the conditional probability that a learned modality-specific representation $Z^{(j)}$ belongs to class a given that the representation from the other modality $Z^{(j')}$ belongs to class y. Each entry $Q^{(j)}[a, y]$ is computed by aggregating the soft assignment values $\phi(Z_i^{(j)}, a)$ over all $i \in [n]$ such that $Y_i = y$, followed by normalization with respect to the total number of samples in that class. The accumulation step (lines 2–7 in Algorithm 3) requires $\Theta(nC)$ time, while the normalization step (lines 8–12 in Algorithm 3) requires $\Theta(C^2)$ time. Since $n \gg C$, the computational complexity of computing $Q^{(j)}$ is $\Theta(nC + C^2) = \Theta(nC)$.

Once $Q^{(j)}(Z^{(j)} \in a \mid Z^{(j')} \in y)$ is computed for $a \in [C]$ and $y \in [C]$, $V_i^{(j)}$ can be computed by $V_i^{(j)} = \sum_{a=1}^{C} \sum_{y=1}^{C} \phi\Big(Z_i^{(j)} \in a,\, Z_i^{(j')} \in y\Big) \log Q^{(j)}(Z^{(j)} \in a \mid Z^{(j')} \in y)$, which takes $\Theta(C^2)$ time for each $i \in [n]$ As a result, the computation of $\frac{1}{n} \sum_{i=1}^{n} V_i^{(j)}$ takes $\Theta(nC^2)$ time. Since $\text{IBB}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} U_i^{(j)} - \frac{1}{n} \sum_{i=1}^{n} V_i^{(j)}$, the overall computation cost for $\text{IBB}^{(j)}$ is $\Theta(nT_0 + nC + nC + nC + nC^2) = \Theta(nT_0 + nC^2)$.