RedCore: Relative Advantage Aware Cross-Modal Representation Learning for Missing Modalities with Imbalanced Missing Rates

Jun Sun[1], Xinxin Zhang[2], Shoukang Han[1], Yu-Ping Ruan[1], Taihao Li[1]

[1]Institute of Artificial Intelligence, Zhejiang Lab, Hangzhou, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
sunjun16sj@gmail.com, zhangxinxin221@mails.ucas.ac.cn, {hanshoukang, ypruan, lith}@zhejianglab.com

Abstract

Multimodal learning is susceptible to modality missing, which poses a major obstacle for its practical applications and, thus, invigorates increasing research interest. In this paper, we investigate two challenging problems: 1) when modality missing exists in the training data, how to exploit the incomplete samples while guaranteeing that they are properly supervised? 2) when the missing rates of different modalities vary, causing or exacerbating the imbalance among modalities, how to address the imbalance and ensure all modalities are well-trained? To tackle these two challenges, we first introduce the variational information bottleneck (VIB) method for the cross-modal representation learning of missing modalities, which capitalizes on the available modalities and the labels as supervision. Then, accounting for the imbalanced missing rates, we define relative advantage to quantify the advantage of each modality over others. Accordingly, a bi-level optimization problem is formulated to adaptively regulate the supervision of all modalities during training. As a whole, the proposed approach features **Re**lative **ad**vantage aware **C**ross-m**o**dal **r**epresentation le**a**rning (abbreviated as **RedCore**) for missing modalities with imbalanced missing rates. Extensive empirical results demonstrate that RedCore outperforms competing models in that it exhibits superior robustness against either large or imbalanced missing rates. The code is available at: https://github.com/sunjunaimer/RedCore.

Introduction

Multimodal learning leverages heterogeneous and comprehensive signals, such as acoustic, visual, lexical information to perform typical machine learning tasks, for instance, clustering, regression, classification, and retrieval (Sun, Dong, and Liu 2021; Zhang et al. 2022; Han et al. 2022). Relative to its unimodal counterpart, multimodal learning has demonstrated great success in numerous applications, including but not limited to medical analysis (Liu et al. 2023), action recognition (Woo et al. 2023), affective computing(Sun et al. 2023). Nevertheless, multimodal learning is inevitably faced with the modality missing issue due to malfunctioning sensors, high data acquisition costs, privacy concerns, etc. In this circumstance, a model trained with full modalities can be highly sensitive to modality missing in inference

phase, and thus suffers from considerable performance deterioration in practical scenarios where some modalities are unavailable. In order to guarantee the reliability and safety in practice, to develop multimodal models robust against modality missing has raised widespread interest from industry and research communities (Tang and Liu 2022).

In the presence of missing modalities, popular studies usually first conduct missing modality imputation before further computations. Naive zero or mean value (Wen et al. 2021; Hu and Chen 2018) padding is of minimal exertion, yet can be quite inaccurate. More advanced statistical techniques primarily rely on the nonparametric Bayesian approach (Manrique-Vallier and Reiter 2017), k-nearest neighbors (Jadhav, Pramod, and Ramanathan 2019), Gaussian processes (Jafrasteh et al. 2023). Current prevalent models building upon deep learning are more capable of exploring the correlation between modalities to impute the missing ones. It is straightforward that generative methods like generative adversarial network (Yoon and Sull 2020) and variational autoencoder (Rezaei-Shoshtari et al. 2021), can be applied to the reconstruction of missing modalities. Besides, knowledge distillation can also be employed to predict the information of missing modalities from a large model pretrained with large corpora of modality-complete data (Wang et al. 2020; Wei, Luo, and Luo 2023).

According to whether all collected data is utilized for training, current schemes can be categorized into two regimes: 1) perfect data training (PDT) (Zhao, Li, and Jin 2021; Woo et al. 2023), where samples with missing modalities are discarded, and only the complete samples are used for model training; and 2) imperfect data training (IDT)(Ma et al. 2021), where samples with missing modalities are retained for training together with complete samples. In PDT, some modalities of the training samples are masked to simulate the absence of modalities, which enforces the model to be aware of the situation of modality missing; although the masked modalities do not contribute directly to the forward feature extraction, they play a role in supervising the learning procedure, as illustrated in Figure 1. In IDT, albeit the missing modalities of imperfect samples are absolutely inaccessible, their available modalities carrying useful information can facilitate the model training. In comparison, the obvious merit of IDT lies in that it is more data-efficient than PDT, as all collected data is fully exploited. However,
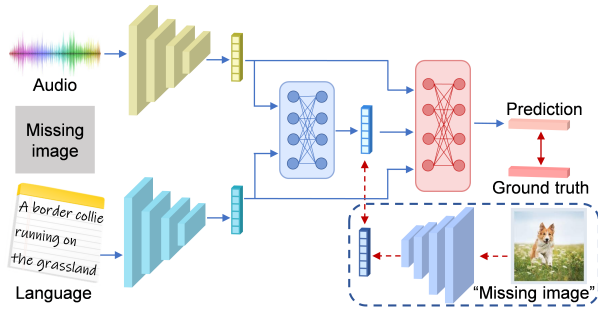
Figure 1: The typical architecture of PDT. With the image module in the dashed blue box removed, it corresponds to IDT. The bi-directional arrows represent supervision.

the downside is that the model design and training of IDT is more challenging, since the lack of supervision for the imputation can negatively impact the model. To sum up, the dilemma of prior works is that they have to compromised between data efficiency and supervision sufficiency, which is the first motivation of the present work.

Furthermore, in practical scenarios, the missing rates of different modalities can vary significantly. For instance, it is likely that 80% of all training samples might miss one modality, while only 20% might miss another modality. Consequently , this can lead to modality imbalance, where weak modalities (those with high missing rates in our case) might be dominated by strong ones, suffering from being under-trained. Modality imbalance problem is inherent for multimodal learning, since some modalities can naturally perform the task better and learn faster than others. Despite the efforts of previous research to address the modality imbalance issue (Sun, Mai, and Hu 2021; Peng et al. 2022; Wang, Tran, and Feiszli 2020; Wu et al. 2022), it requires special attention in the context of modality missing. There are two reasons: firstly, modality imbalance can be aggravated by imbalanced missing rates; secondly, apart from the final task, imputation is also influenced by imbalance missing rates. To the best of our knowledge, no existing relevant study has taken the imbalanced missing rates across modalities into account, which constitutes our second motivation.

With the above analysis, our work aims to achieve two primary goals: 1) to develop a model that enjoys the data efficiency and supervision sufficiency simultaneously for multimodal learning with modality missing; and 2) to propose a training scheme to solve the problem of imbalance missing rates. For the model design and training, we adopt the IDT regime, which, as mentioned above, is data efficient. Taking advantage of the variational information bottleneck (VIB) method (Alemi et al. 2016; Bang et al. 2021), we devise a model with cross-modal representation learning for missing modality imputation. With VIB, each modality learns an optimal representation that retains minimal information from the original feature while capturing the maximal information relevant to the task. The novelty of our method is that for each modality, regardless of being missing or available, a cross-modal representation generated from the optimal rep-

resentation of all other available modalities is learned. The supervision of the cross-modal representation learning for available modalities is from both the features and the labels, and for missing modalities it comes from only the labels.

In order to balance the modalities, we first define the relative advantage of each individual modality according to the imputation loss during training. Subsequently, with relative advantage, a bi-level optimization problem (Ji 2021; Chen et al. 2023) is formulated to regulate the supervisions of different modalities. Based on the analysis of the bi-level optimization problem we propose a two-double algorithm, corresponding to the joint training and adaptive supervision regulation approach.

In summary, we propose a novel approach — **Re**lative a**d**vantage aware **C**ross-m**o**dal **r**epresentation l**e**arning, referred to as **RedCore**, which addresses the challenge of missing modalities with imbalanced missing rates. The contributions are primarily threefold:

- Building upon the VIB method, we leverage the available modalities and the labels as supervision for cross-modal representation learning, guaranteeing that the imputation modules are sufficiently supervised in the data-efficient IDT training framework.

- We propose the concept of relative advantage for modalities, according to which a bi-level optimization problem is formulated to adaptively balance their training.

- Numerical results validate the effectiveness of RedCore and demonstrate its superior performance compared with state-of-the-art models.

Related Works
Missing Modality Imputation
Imputation is of central importance in the literature addressing modality missing. The vast majority of existing works fall in the category of implicit imputation (discriminative approach) and explicit imputation (generative approach) (Morales-Alvarez et al. 2022; Miyaguchi et al. 2022). The former carries out imputation implicitly or simultaneously with the prediction (Josse et al. 2020; Jeong, Wang, and Calmon 2022). The latter typically maintains an explicit generative module for missing value completion, which spans a wide spectrum of schemes, including kernel learning, graph learning, matrix factorization, and subspace learning. Kernel learning is usually adopted in clustering task with incomplete modalities, where the clustering and kernel feature learning can be performed alternatively or simultaneously (Liu et al. 2018, 2020). For graph based approaches, subgraph of each accessible modality is usually pre-constructed, and cross-modal graph learning is proposed to capture the missing local graph structure and complete the global graph (Li et al. 2022). Matrix factorization can be utilized to obtain the latent factors of missing modalities by leveraging complementary multimodal information and underlying crossmodal relation (Liu et al. 2021). Subspace learning maps the multimodal features into low-dimensional latent representations, and exploit the correlation between modalities to impute the missing latent representation. A unified subspace embedding alignment framework is proposed in (Liu

et al. 2020), where the missing modalities are reconstructed with Laplacian regularized error matrices compensation. Besides, autoencoder based methods are representative subspace learning approaches for missing modality generation (Du et al. 2018). Via stacking a series of residual autoencoders, cascaded residual autoencoder (CRA) is designed to promote the missing modality modeling ability of autoencoders (Tran et al. 2017). Additionally, incorporating cycle consistency loss to CRA can further improve the performance of cross-modal imputation (Zhao, Li, and Jin 2021).

Fairness and Imbalance in Multimodal Learning

Because of the modality discrepancy, multimodal learning naturally encounters the concern of fairness and imbalance (Zhang and Long 2021), which is highlighted in context of modality incompleteness. The fairness of machine learning methods with missing attributes are comprehensively investigated in (Fernando et al. 2021). A reweighting scheme that assigns lower weight to data sample with missing values to promote fairness is proposed in (Wang and Singh 2021). A decision tree without explicit imputation is trained by optimizing a fairness-regularized objective function in (Jeong, Wang, and Calmon 2022). Another line of works that inspire our proposed approach to addressing the modality imbalance are those coping with the imbalanced convergence rates of different modalities in multimodal model training. The learning rates of different modalities are adaptively tuned during training according to the loss, the overfitting estimation, and the parameters & their gradients of each modality in works (Sun, Mai, and Hu 2021; Peng et al. 2022), (Wang, Tran, and Feiszli 2020) and (Wu et al. 2022), respectively. Our work distinguishes from these prior studies in two aspects: 1) we particularly focus on balancing the training of representation imputation; and 2) in stead of balancing the modalities heuristically, we formulate an optimization problem for supervision regulation.

Method: RedCore

The proposed approach, RedCore, is composed of two components — Core and Red, as elaborated in the following subsections. Prior to that, we give some notation definitions.

Notations: Suppose a multimodal training dataset contains N samples, each with M modalities. For ease of expression, let us define an auxiliary modality as a union of all modalities, and thus the total number of modalities is $M + 1$. Let $[P]$ for any positive integer P denote the set $\{1, 2, \cdots, P\}$. The training samples are denoted by $(\{\boldsymbol{x}_n^m\}_{m\in[M]}, \{y_n^m\}_{m\in[M+1]})$, where $n \in [N]$ indexes the samples, $\boldsymbol{x}_n^m \in \mathbb{R}^{d_m}$ represents the d_m-dimensional feature vector (the feature can also be vector sequence) of modality $m, \forall m \in [M]$, and y_n^m represents the label corresponding to modality $m, \forall m \in [M + 1]$ (for datasets where all modalities share a common label, $y_n^1 = y_n^2 = \cdots = y_n^{M+1}$ holds). For the consistency of expression, we use $\boldsymbol{x}_n^{M+1} := [\boldsymbol{x}_n^1; \boldsymbol{x}_n^2; \cdots; \boldsymbol{x}_n^M]$ to collect all features of sample n. Let matrix $\boldsymbol{C} \in \mathbb{R}^{N \times M}$ be the modality presence indication matrix, and let C_{nm} denote the n, m-th entry of $\boldsymbol{C}$. Then, $C_{nm} = 1$, if modality m of sample n is available; otherwise,

$C_{nm} = 0$. For missing modalities, their labels are endowed with the joint label; namely, for any sample $n, n \in [N]$ with modality missing, $y_n^m = y_n^{M+1}$ holds for all $m \in \{m \mid C_{nm} = 0\}$. The missing rate of modality $m, \forall m \in [M]$, MR^m, is calculated as $MR^m = (N - \sum_{n\in[N]} C_{nm})/N$. We assume that $MR^m \in [0, 1)$ to ensure that all modalities exist in the training dataset.

Let $\boldsymbol{x}^m$ and $y^m, \forall m \in [M + 1]$ respectively denote the general random variables of the feature and label, and $\boldsymbol{x}_n^m$ and y_n^m are realizations of them, respectively. Let $\boldsymbol{z}^m \in \mathbb{R}^d$, a map of $\boldsymbol{x}^m$, denote the representation of modality m. Similarly, $\boldsymbol{z}_n^m$ is a realization of $\boldsymbol{z}^m$ (for brevity, we assume the representations of all modalities are d-dimensional vectors).

Core — Cross Modal Representation Learning

In this subsection, we will introduce the framework of our model and VIB based representation learning.

A. Model Architecture The proposed model architecture is visualized in Figure 2, where Figure 2(a) shows the training computation flow on a sample with 3 complete modalities, Figures 2(b) and 2(c) exhibit the training and inference with modality 2 missing. In specific, the feature of each available modality $\boldsymbol{x}^m$ is mapped to its representation $\bar{\boldsymbol{z}}^m$ via a variational encoder with parameter $\boldsymbol{\theta}_{senc}^m$ (senc: self-modal encoder). Subsequently, through cross-modal encoder with parameter $\boldsymbol{\theta}_{xenc}^m$ (xenc: cross-modal encoder), each modality obtains a representation $\hat{\boldsymbol{z}}^m$ from the representation $\bar{\boldsymbol{z}}^m$ of all other available modalities. The final representation $\boldsymbol{z}^m$ is attained as $\bar{\boldsymbol{z}}^m$ (resp., $\hat{\boldsymbol{z}}^m$) if modality m is present (resp., missing). Mathematically, for feature $\boldsymbol{x}_n^m$, its representation $\boldsymbol{z}_n^m$ is derived as following. Firstly, the representations of available modalities are sampled from distribution $p(\bar{\boldsymbol{z}}^m|\boldsymbol{x}^m; \boldsymbol{\theta}_{senc}^m)$, i.e., $\bar{\boldsymbol{z}}_n^m \sim p(\bar{\boldsymbol{z}}^m|\boldsymbol{x}^m; \boldsymbol{\theta}_{senc}^m), \forall m \in \{m \mid C_{nm} = 1\}$. Then, the cross-modal representations are calculated for all modalities as $\hat{\boldsymbol{z}}_n^m = g(\{\bar{\boldsymbol{z}}_n^{m'}\}_{m'\in\mathcal{A}_n^m}; \boldsymbol{\theta}_{xenc}^m)$, where $\mathcal{A}_n^m$ is defined as the set of all available modalities other than modality m of sample n, namely, $\mathcal{A}_n^m := \{m' \in [M] \mid C_{nm'} = 1, m' \neq m\}$. Finally, with the above notations, the ultimate representation of feature $\boldsymbol{x}_n^m, \forall n \in [N], \forall m \in [M]$ boils down to:

$$\boldsymbol{z}_n^m = \begin{cases} \bar{\boldsymbol{z}}_n^m, & \text{if } C_{nm} = 1; \\ \hat{\boldsymbol{z}}_n^m, & \text{otherwise.} \end{cases} \quad (1)$$

The representation of modality $M + 1$ is the concatenation of z_n^m for all $m \in [M]$, i.e., $\boldsymbol{z}_n^{M+1} := [\boldsymbol{z}_n^1; \boldsymbol{z}_n^2; \cdots; \boldsymbol{z}_n^M]$. When we use $\boldsymbol{\theta}_{enc}^m := \{\boldsymbol{\theta}_{senc}^m, \boldsymbol{\theta}_{xenc}^m\}$ to collect encoder parameters for modality m, and $\boldsymbol{\theta}_{enc}^{M+1} := \{\boldsymbol{\theta}_{enc}^m\}_{m\in[M]}$ to gather all encoder parameters, the above representation can be summarized as $\boldsymbol{z}_n^m \sim p(\boldsymbol{z}^m|\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m), \forall m \in [M + 1]$.

For a typical multimodal learning setting, the joint representation $\boldsymbol{z}_n^{M+1}$ is used for the final task, which is tantamount to passing the representation to a decoder with parameter $\boldsymbol{\theta}_{dec}^{M+1}$ and yielding the label prediction: $\hat{y}_n^{M+1} \sim p(y^{M+1}|\boldsymbol{z}^{M+1}; \boldsymbol{\theta}_{dec}^{M+1})$. In addition, in our framework, each individual modality also performs label prediction, even if all modalities share the same label ($y_n^1 = y_n^2 = \cdots = y_n^{M+1}$). This relies on the decoder $\boldsymbol{\theta}_{dec}^m$: $y_n^m \sim p(y^m|\boldsymbol{z}^m; \boldsymbol{\theta}_{dec}^m), m \in$
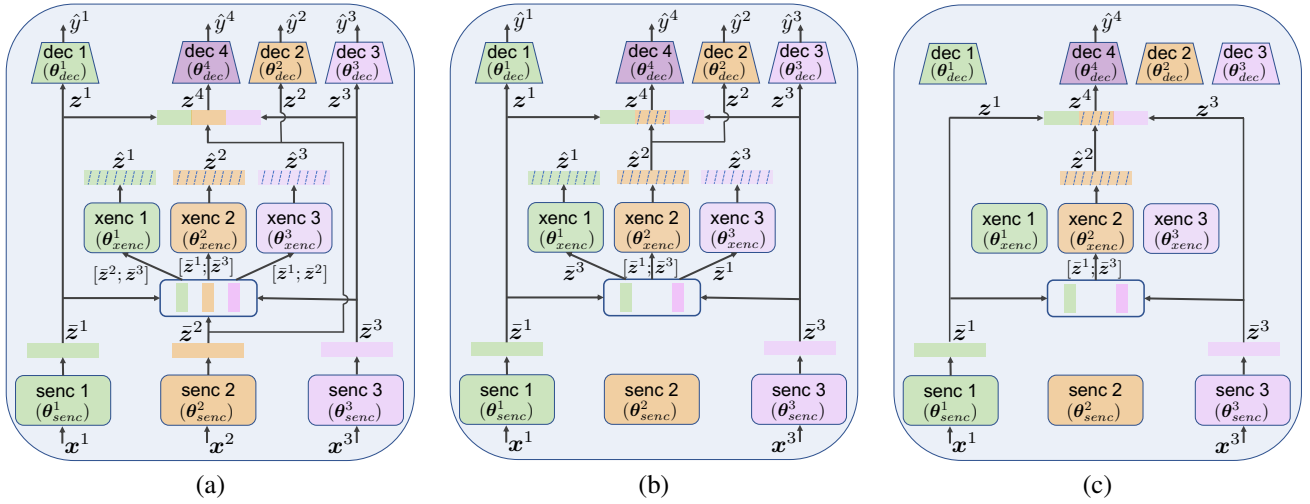
Figure 2: Model architecture (with 3 modalities as an example). (a) Training on a sample with all modalities available. (b) Training on a sample with modality 2 missing. (c) Inference on a sample with modality 2 missing.

$[M]$. The reason why these additional predictions are conducted will be given in the next subsection.

B. VIB Based Representation Learning The goal of variational information bottleneck based learning is to obtain a representation that is maximally informative about the target and maximally compressive about the raw feature. In our scenario, this can be cast as the following optimization problem which minimizes the information bottleneck (IB) loss.

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{IB}(\boldsymbol{\theta}) := \sum_{m \in [M+1]} I(\boldsymbol{x}^m, \boldsymbol{z}^m; \boldsymbol{\theta}) - \gamma I(\boldsymbol{z}^m, y^m; \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} := \{\boldsymbol{\theta}_{enc}^m, \boldsymbol{\theta}_{dec}^m\}_{m \in [M+1]}$ collects all model parameters, $I(\cdot, \cdot)$ represents the mutual information between two random variables, and the coefficient $\gamma > 0$ is a predefined constant that balances the two terms in the loss function.

Since it is intractable to compute the mutual information $I(\boldsymbol{x}^m, \boldsymbol{z}^m; \boldsymbol{\theta})$ and $I(\boldsymbol{z}^m, y^m; \boldsymbol{\theta})$, the variational upper bound of $\mathcal{L}_{IB}(\boldsymbol{\theta})$ is invoked (Alemi et al. 2016):

$$\mathcal{L}_{IB}(\boldsymbol{\theta}) \leq \mathcal{L}_{VIB}(\boldsymbol{\theta})$$
$$:= \sum_{m=1}^{M+1} \mathbb{E}[D_{KL}(p(\boldsymbol{z}^m|\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m)||q(\boldsymbol{z}^m)) \quad (2)$$
$$- \gamma \log q(y^m|\boldsymbol{z}^m; \boldsymbol{\theta}_{dec}^m)],$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to random variables $\boldsymbol{x}^m, \boldsymbol{z}^m, y^m$; $D_{KL}(\cdot||\cdot)$ denotes the KL-divergence between two distributions; $q(\boldsymbol{z}^m)$ and $q(y^m|\boldsymbol{z}^m; \boldsymbol{\theta}_{dec}^m)$ denote variational distributions approximating $p(\boldsymbol{z}^m|\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m)$ and $p(y^m|\boldsymbol{z}^m; \boldsymbol{\theta}_{dec}^m)$, respectively.

The variational encoders take the Gaussian form $p(\boldsymbol{z}^m|\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m) = \mathcal{N}(\boldsymbol{z}^m; f_{\boldsymbol{\mu}}(\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m), f_{\boldsymbol{\Sigma}}(\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m))$, where $f_{\boldsymbol{\mu}}(\cdot; \boldsymbol{\theta}_{enc}^m) \in \mathbb{R}^d$ and $f_{\boldsymbol{\Sigma}}(\cdot; \boldsymbol{\theta}_{enc}^m) \in \mathbb{R}^{d \times d}$ are the mean vector and covariance matrix determined by the network models as will be specified in the Numerical Results

section. $q(\boldsymbol{z}^m)$ is fixed as a d-dimensional spherical Gaussian, that is $q(\boldsymbol{z}^m) = \mathcal{N}(\boldsymbol{z}^m; \mathbf{0}_d, \boldsymbol{I}_d)$, where $\mathbf{0}_d$ and $\boldsymbol{I}_d$ are d-dimensional all-zero vector and $d \times d$ identity matrix, respectively. The decoders are simple classifier models of the form $q(y^m|\boldsymbol{z}^m) = \text{softmax}(f_c(\boldsymbol{z}^m))$, where $f_c(\boldsymbol{z}^m), m \in [M+1]$ are MLPs that map the representations to the logits for classification. Applying the reparameterization trick (Kingma and Welling 2014), we obtain $\boldsymbol{z}^m$ as follows:

$$\boldsymbol{z}^m = f_{\boldsymbol{\mu}}(\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m) + \boldsymbol{\epsilon} \odot f_{\boldsymbol{\Sigma}}(\boldsymbol{x}^m; \boldsymbol{\theta}_{enc}^m)),$$

where $\odot$ means element-wise product and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$. As a result, $\mathcal{L}_{VIB}(\boldsymbol{\theta})$ can be approximated by the following $\tilde{\mathcal{L}}_{VIB}(\boldsymbol{\theta})$:

$$\tilde{\mathcal{L}}_{VIB}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M+1} D_{KL}(p(\boldsymbol{z}_n^m|\boldsymbol{x}_n^m; \boldsymbol{\theta}_{enc}^m)||q(\boldsymbol{z}_n^m)) \quad (3)$$
$$- \gamma \mathbb{E}_{\boldsymbol{\epsilon}}[\log q(y_n^m|\boldsymbol{z}_n^m; \boldsymbol{\theta}_{dec}^m)],$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ is the random variable from the reparameterization trick. This formulation circumvents the expectation in eq.(2), and thus allows us to compute unbiased stochastic gradient via backpropagating of a single sample.

Minimizing $\tilde{\mathcal{L}}_{VIB}(\boldsymbol{\theta})$ leads to an optimal representation for each modality in the sense of information bottleneck theory, meaning compressing the feature and meanwhile maximizing the prediction accuracy, which corresponds respectively to the first and second terms in eq. (3). This explains why we enforce each modality to predict its own label — to attain its optimal task-relevant representation. Consequently, the cross-modal learning module is freed from dealing with task-irrelevant details, which thereby relieves the difficulty in learning parameters $\boldsymbol{\theta}_{xenc}^m, \forall m \in [M]$.

C. Cross-modal Representation Learning The imputation in our framework translates to the cross-modal representation, which is learned for all modalities regardless of

whether they are present or absent for each sample. For both missing and available modalities, the cross-modal encoder receives supervision from the label, which can be achieved via minimizing the VIB loss $\tilde{\mathcal{L}}_{VIB}(\boldsymbol{\theta})$. For available modalities, the cross-modal encoder receives additional supervision from the feature by minimizing the following mean squared error (MSE), i.e., the imputation loss:

$$\mathcal{L}_{MSE}^m(\boldsymbol{\theta}) := \frac{1}{N_m} \sum_{n=1}^N ||\boldsymbol{z}_n^m - \hat{\boldsymbol{z}}_n^m||_2^2,$$

where $N_m = \sum_{n\in[N]} C_{nm}$ is the number of samples that have modality m. From the definition of $\boldsymbol{z}_n^m$ in eq. (1), it can be observed that the imputation loss $\mathcal{L}_{MSE}^m(\boldsymbol{\theta})$ accounts only for the representation recovery loss of available modalities (because $\boldsymbol{z}_n^m = \hat{\boldsymbol{z}}_n^m$ holds for missing modalities).

The overall loss function for training is

$$\mathcal{L}(\boldsymbol{\theta}) := \tilde{\mathcal{L}}_{VIB}(\boldsymbol{\theta}) + \sum_{m\in[M]} \eta^m \mathcal{L}_{MSE}^m(\boldsymbol{\theta}),$$

where the coefficients $\eta^m, \forall m \in [M]$ weigh the representation reconstruction losses of different modalities.

To summarize, the supervisions of the encoders, including both the self-modal representation encoder ($\boldsymbol{\theta}_{senc}^m, m \in [M]$) and the cross-modal representation encoder ($\boldsymbol{\theta}_{xenc}^m, m \in [M]$), come from the label and the available modalities. As Figure 2(b) illustrates, when training on a certain sample with missing modality m, the parameter $\boldsymbol{\theta}_{senc}^m$ and $\boldsymbol{\theta}_{xenc}^m$ do not receive supervision from the feature, yet the latter are supervised by the label. Therefore, in the circumstance of imbalanced missing rates, the parameters corresponding to modalities with larger missing rates receive supervision less frequently. This can cause some modalities to be overfitting while others underfitting, which is the issue to be addressed in the following subsection.

Red — Relative Advantage Aware Supervision Regulation

In this part, we start with defining the relative advantage of each modality to quantify how strong (or weak) it is compared to other modalities. Then, according to the relative advantage, the strength of supervision, η^m, is adaptively regulated to balance the learning process across all modalities.

A. Relative Advantage   The strong modalities are usually better at learning or have smaller missing rates, which effects relatively smaller imputation losses during training. This inspires us to define the relative advantage according to the imputation loss. Concretely, for any $\boldsymbol{\theta}$, the average imputation loss over modalities is first calculated as:

$$\bar{\mathcal{L}}_{MSE}(\boldsymbol{\theta}) = \sum_{m\in[M]} \mathcal{L}_{MSE}^m(\boldsymbol{\theta})/M. \qquad (4)$$

With $\bar{\mathcal{L}}_{MSE}^m(\boldsymbol{\theta})$, relative advantage is formally defined as:

$$RA^m(\boldsymbol{\theta}) := [\bar{\mathcal{L}}_{MSE}(\boldsymbol{\theta}) - \mathcal{L}_{MSE}^m(\boldsymbol{\theta})]/\bar{\mathcal{L}}_{MSE}(\boldsymbol{\theta}). \qquad (5)$$

When $RA^m(\boldsymbol{\theta}) > 0$ (resp., < 0), modality m is considered strong (resp., weak) relative to other modalities; $RA^m(\boldsymbol{\theta}) = 0$ indicates that modality m is at the balancing point.

B. Relative Advantage Aware Supervision Regulation
For brevity, let $\boldsymbol{\eta} = [\eta^1, \eta^2, \cdots, \eta^M]^T$ and $\boldsymbol{r}(\boldsymbol{\theta}) = [RA^1(\boldsymbol{\theta}), RA^2(\boldsymbol{\theta}), \cdots, RA^M(\boldsymbol{\theta})]^T$ collect all the supervisions and relative advantages, respectively. In order to balance the modalities, the strength of the supervision $\boldsymbol{\eta}$ should be negatively correlated with the relative advantage $\boldsymbol{r}(\boldsymbol{\theta})$. Therefore, the objective of supervision regulation is to minimize the correlation between $\boldsymbol{\eta}$ and $\boldsymbol{r}(\boldsymbol{\theta})$. The joint training and supervision regulation problem can be formulated as the following bi-level optimization problem for the non-trivial case where $\boldsymbol{r}(\boldsymbol{\theta}) \neq \boldsymbol{0}$ ($\boldsymbol{0}$ is an all-zero vector of proper size):

$$\min_{\boldsymbol{\eta},\boldsymbol{\theta}} \quad \boldsymbol{r}(\boldsymbol{\theta})^T \boldsymbol{\eta} \qquad (6a)$$

$$\text{s.t.} \quad \boldsymbol{\eta} \geq \xi_1 \boldsymbol{1} \qquad (6b)$$

$$||\boldsymbol{\eta}||_p = \xi_2 \qquad (6c)$$

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{VIB}(\boldsymbol{\theta}) + \sum_{m\in[M]} \eta^m \mathcal{L}_{MSE}^m(\boldsymbol{\theta}) \qquad (6d)$$

where $\xi_1 > 0$ is a small constant to guarantee that the supervision is positive ($\boldsymbol{1}$ is an all-one vector); constraint (6c) constrains the ℓ_p-norm ($p \geq 1$) of the supervision by a given constant ξ_2, and guarantees the stability of the training; constraint (6d) which involves the model training given the supervision is the lower level problem embedded in the upper one, eq. (6). The upper level problem, with optimization variable $\boldsymbol{\eta}$, corresponds to the supervision regulation — adaptively tuning the supervision via minimizing the correlation between the supervision and the relative advantage.

In the sequel, we will analyze the property of the above problem, and propose our algorithm accordingly. A bi-level optimization problem is generally non-convex, which applies to problem (6). Nonetheless, an appealing characteristic of problem (6) holds, as is stated below.

Lemma 1. *For any fixed parameter $\boldsymbol{\theta}$, problem (6) is equivalent to the following convex problem.*

$$\min_{\boldsymbol{\eta}} \quad \boldsymbol{r}(\boldsymbol{\theta})^T \boldsymbol{\eta} \qquad (7a)$$

$$\textit{s.t.} \quad \boldsymbol{\eta} \geq \xi_1 \boldsymbol{1} \qquad (7b)$$

$$||\boldsymbol{\eta}||_p \leq \xi_2 \qquad (7c)$$

Proof. It follows from the definition of $\boldsymbol{r}(\boldsymbol{\theta})$ that:

$$\boldsymbol{r}(\boldsymbol{\theta})^T \boldsymbol{1} = 0. \qquad (8)$$

Thus, for non-trial case, i.e., $\boldsymbol{r}(\boldsymbol{\theta}) \neq \boldsymbol{0}$, there exists an index $i \in [M]$, such that $RA^i(\boldsymbol{\theta}) < 0$.

We define $\Gamma_1 = \{\boldsymbol{\eta} \in \mathbb{R}^M \mid \boldsymbol{\eta} \geq \xi_1\boldsymbol{1}, ||\boldsymbol{\eta}||_p = \xi_2\}$ and $\Gamma_2 = \{\boldsymbol{\eta} \in \mathbb{R}^M \mid \boldsymbol{\eta} \geq \xi_1\boldsymbol{1}, ||\boldsymbol{\eta}||_p \leq \xi_2\}$ to denote the feasible domains of problems (6) and (7) with $\boldsymbol{\theta}$ fixed, respectively. It is evident that Γ_1 is non-convex due to the nonlinear equality constraint (6c), and Γ_2 is convex. Lemma 1 can be interpreted as the optimal solution to problem (7) lies in Γ_1, which will be proved by contradiction in the following.

Assume $\boldsymbol{\eta}^* \in \Gamma_2$ is an optimal solution to problem (7) and $\boldsymbol{\eta}^* \notin \Gamma_1$, which means $\boldsymbol{\eta}^*$ is an inner point of Γ_2. Consequently, there exists a sufficiently small positive constant $\delta > 0$ such that $\boldsymbol{\eta}' = \boldsymbol{\eta}^* + \delta \cdot \boldsymbol{1}_i \in \Gamma_2$, where

$\mathbf{1}_i$ denotes a vector in which the i-th element is 1 and all other elements are 0. Given that $RA^i(\boldsymbol{\theta}) < 0$, we have $\boldsymbol{r}(\boldsymbol{\theta})^T \boldsymbol{\eta}' < \boldsymbol{r}(\boldsymbol{\theta})^T \boldsymbol{\eta}^*$, indicating that $\boldsymbol{\theta}^*$ is not the optimal solution. This contradicts the initial assumption and concludes the proof. □

For any fixed parameter $\boldsymbol{\theta}$, it is obvious that constraint (6d) can be directly eliminated. Then, the problem (7) is an immediate result of relaxing the nonlinear equality constraint (6c) to inequality (7c). With Lemma 7, for a fixed parameter $\boldsymbol{\theta}$, we apply projected gradient descent to update $\boldsymbol{\eta}$ as the following eq. (9), for which the convergence is guaranteed (Vu and Raich 2022).

$$\boldsymbol{\eta}_{s+1} = Pro_{\Gamma_1}[\boldsymbol{\eta}_s - \alpha_1 \boldsymbol{r}(\boldsymbol{\theta}_s)], \tag{9}$$

where s indexes the updating step of $\boldsymbol{\eta}$, $\Gamma_1 = \{\boldsymbol{\eta} \in \mathbb{R}^M \mid \boldsymbol{\eta} \geq \xi_1 \mathbf{1}, \|\boldsymbol{\eta}\|_p = \xi_2\}$ is the feasible domain of $\boldsymbol{\eta}$, $Pro_{\Gamma_1}[\cdot]$ represents projection to domain Γ_1, and $\boldsymbol{r}(\boldsymbol{\theta}_s)$ is the gradient of the objective function $\boldsymbol{r}(\boldsymbol{\theta}_s)^T \boldsymbol{\eta}$ w.r.t. $\boldsymbol{\eta}$. It is straightforward from eq. (9) that the supervision regulation amounts to reducing the strength of its supervision if modality $m (m \in [M])$ has advantage over others, i.e., $RA^m(\boldsymbol{\theta}) > 0$; otherwise, the supervision strength should be increased.

C. Training with Dynamic Supervision Regulation The optimizer for training is usually based on stochastic gradient, and consequently, the loss might fluctuate significantly with iterations. To alleviate the fluctuation, we use the following moving average loss (MAL) to replace $\mathcal{L}^m_{MSE}(\boldsymbol{\theta})$ in eqs. (4) and (5) for relative advantage calculation.

$$\mathcal{L}^m_{MAL}(\boldsymbol{\theta}_k) = (1 - \beta)\mathcal{L}^m_{MAL}(\boldsymbol{\theta}_{k-1}) + \beta \mathcal{L}^m_{MSE}(\boldsymbol{\theta}_k), \tag{10}$$

where k represents the index of $\boldsymbol{\theta}$ updating step, and $\beta \in (0, 1]$ is a pre-specified constant. Following the work (Yang, Ji, and Liang 2021), we propose a double-loop algorithm to solve the bi-level optimization problem, as summarized in Algorithm 1. The supervision $\boldsymbol{\eta}$ is iteratively updated for every K-step updates of the model parameter $\boldsymbol{\theta}$.

Numerical Results

Benchmark Datasets and Models: In this section we evaluate RedCore on three commonly used multimodal datasets: IEMOCAP (Busso et al. 2008), CMU-MOSEI (Zadeh et al. 2018) and MSP-IMPROV (Busso et al. 2016). All these datasets include three modalities: acoustic (A), visual (V) and lexical (L). IEMOCAP contains conversation videos where two actors perform improvised or scripted scenarios. The conversations are divided into utterances which are annotated into categorical emotion labels. Following the popular label processing method as in work (Liang, Li, and Jin 2020), we obtain the four-class dataset. MSP-IMPROV consists of data from six dyadic conversations involving 12 actors. For a fair comparison, we adopt the same video selection scheme described in work (Zhao, Li, and Jin 2021) to set up the four-class emotion recognition dataset. MOSEI is a collection of over 23,500 sentence utterance videos from more than 1000 online YouTube speakers. Each utterance is annotated with sentiment labels: positive, negative and

---

Algorithm 1: Dynamic supervision regulation

1: **Initialization:** randomly choose $\boldsymbol{\eta}_0$ and $\boldsymbol{\theta}_0$.
2: **for** $s = 0$ to $S - 1$ **do**
3:     calculate $\boldsymbol{r}(\boldsymbol{\theta}_{sK})$ according to eqs. (10), (4) and (5);
4:     update $\boldsymbol{\eta}$ according to eq. (9) with $\boldsymbol{r}(\boldsymbol{\theta}_{sK})$.
5:     **for** $k = 0$ to $K - 1$ **do**
6:         update $\boldsymbol{\theta}$ using stochastic gradient descent (or other schemes, e.g., Adam):
7:         $\boldsymbol{\theta}_{sK+k+1} = \boldsymbol{\theta}_{sK+k} - \alpha_2 \tilde{\nabla}\mathcal{L}(\boldsymbol{\theta}_{sK+k})$, where $\tilde{\nabla}\mathcal{L}(\boldsymbol{\theta}_{sK+k})$ is the stochastic gradient of $\mathcal{L}(\boldsymbol{\theta}_{sK+k})$.
8:     **end for**
9: **end for**
10: **Return:** Model parameter $\boldsymbol{\theta}_{SK}$.

---

neutral. For datasets IEMOCAP and MSP-IMPROV, feature extraction method in (Zhao, Li, and Jin 2021) is employed to obtain sequential feature vectors for each modality. We utilize toolkit in work (Mao et al. 2022) to attain feature vectors for CMU-MOSEI. In the comparison studies, the baseline models are cascaded residual autoencoder (CRA) (Zhao, Li, and Jin 2021), action masked autoencoder (AMAE) (Woo et al. 2023) and missing modality imagination network (MMIN) (Zhao, Li, and Jin 2021).

Implementation Details

We employ transformer encoders as self-model encoders (senc) for all modalities, residual autoencoders as cross-modal encoders (xenc) and multilayer perceptrons as decoders (dec). The number of self-attention layers for all self-modal encoders (senc) is 3, and all layers are with 8 heads and 128-dimensional hidden embedding. The cross-modal encoders for all modalities consist of 5 residual autoencoder blocks with RA-layers of size 256-128-64-32-64-128-256, and the outputs are 128-dimensional vectors as the representation for corresponding modalities. The coefficient γ for the VIB loss is set to be 0.008. Optimizer Adam with learning rate 2×10^{-4}, momentum coefficient $(0.9, 0.999)$ and batch size 128 is adopted for model training. The parameter settings for supervision regulation are $p = 2, \xi_1 = 0.015, \xi_2 = 0.15, K = 3, \alpha_1 = 0.1, \beta = 0.7$. More details of the implementation can be found from the code in the supplementary materials[1]. To mimic modality missing, the modalities are randomly masked according to their missing rates. The competing models adopt perfect data training (PDT), meaning the "missing data" is used indirectly for training. In contrast, RedCore applies imperfect data training (IDT), i.e., the missing data is inaccessible. The experiments are carried out on 4 Nvidia Tesla V100 GPUs with 32G memory. Throughout this section, we use weighted F1 score as the performance metric. The reported results are obtained by averaging over 3 repeated experiments.
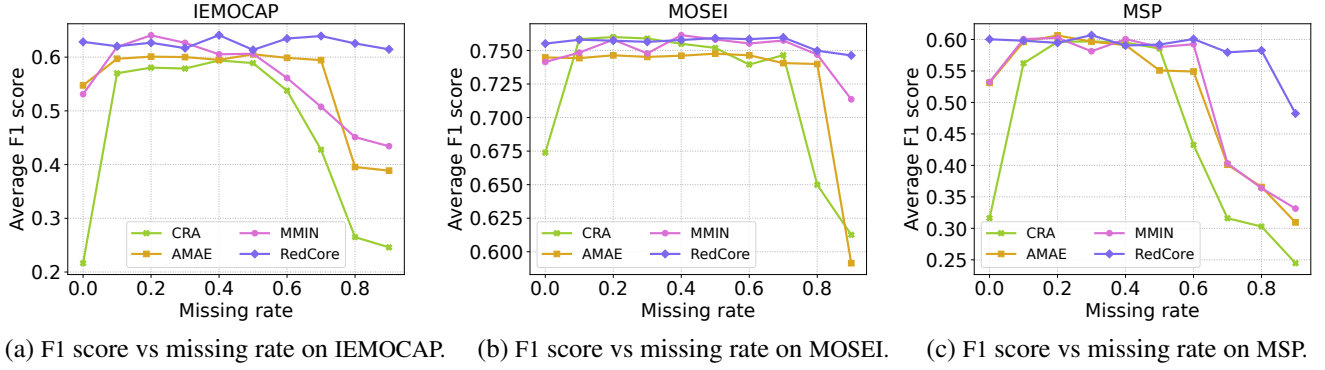
---

[1]https://arxiv.org/abs/2312.10386

(a) F1 score vs missing rate on IEMOCAP.　　(b) F1 score vs missing rate on MOSEI.　　(c) F1 score vs missing rate on MSP.

Figure 3: Average F1 score with varying missing rates for different datasets.

| MR of | (0.8, 0.5, 0.5) | | (0.5, 0.8, 0.5) | | (0.5, 0.5, 0.8) | | (0.2, 0.5, 0.8) | | (0.5, 0.2, 0.8) | | (0.8, 0.5, 0.2) | |
(A, V, L)	M.N.	R.C.	M.N.	R.C.	M.N	R.C.	M.N.	R.C.	M.N.	R.C.	M.N.	R.C.
A	0.5067	0.5129	0.5505	0.5476	0.5054	0.4922	0.5260	0.5122	0.4691	0.5033	0.4842	0.4878
V	0.5746	0.5199	0.5965	0.4520	0.5542	0.5206	0.5771	0.5067	0.6077	0.4995	0.5554	0.4804
L	0.5792	0.6086	0.4639	0.5841	0.2239	0.5879	0.1223	0.5886	0.1499	0.5905	0.2619	0.5769
AV	0.6534	0.6390	0.6026	0.6062	0.5603	0.6479	0.5787	0.5752	0.634	0.6209	0.6073	0.6319
AL	0.6142	0.6654	0.5643	0.6454	0.3973	0.6185	0.5201	0.5897	0.2943	0.6462	0.3102	0.6413
VL	0.6453	0.6867	0.6567	0.6334	0.2283	0.6669	0.5577	0.6573	0.1673	0.6500	0.3020	0.6371
AVL	0.6328	0.7285	0.6343	0.6768	0.4230	0.7051	0.5718	0.6781	0.3629	0.6977	0.3257	0.6961
Ave.	0.6088	**0.6306**	0.5986	**0.6088**	0.4748	**0.618**	0.5347	**0.5999**	0.4514	**0.6109**	0.4393	**0.6017**

Table 1: F1 score of RedCore (R.C.) and MMIN (M.N.) with imbalanced missing rates (MR) for modalities (A, V, L).

Comparison Studies

In the comparative analysis, we evaluate the models with two settings — the training data with the same missing rate for all modalities, and with imbalanced missing rates.

Missing Modalities with the Same Missing Rate: The models are trained on three datasets with shared missing rates ranging from 0.0 to 0.9, and are tested with all combinations of input modalities (i.e., A, V, L, AV, AL, VL, AVL). Detailed results of the F1 score for each input combination are reported by the table in the supplementary materials. Figure 3 illustrates the average F1 score (average over all the input combinations) on the three datasets. Three observations about all tested datasets from Figure 3 demonstrate the superiority of RedCore over the competing models. 1) When the missing rate is low, i.e., 0.0 and 0.1, RedCore outperforms all the baseline models. This is because other compared models are not fully exposed to modality missing, and hence have difficulty in coping with the situation, while RedCore consistently conducts cross-modal representation learning even when no modality is missing. 2) when the missing rate is at a moderate level, (e.g., between 0.2 and 0.6 on IEMOCAP), RedCore shows at least comparable performance with other models, despite adopting IDT training regime with less data. This verifies that RedCore is more data-efficient than the baseline models. 3) when the missing rate is large, (e.g., between 0.7 and 0.9 on IEMOCAP), RedCore can still maintain the performance or have slight to moderate performance drop, whereas the compared mod-

els all show significant performance degradation. This observation validates the exploitation of supervision from the available modalities and the labels boosts the performance of RedCore. In conclusion, RedCore has demonstrated superior robustness over a wide spectrum of missing rates.

Missing Modalities with Imbalanced Missing Rates: Table 1 compares MMIN and RedCore trained with imbalanced missing rates and reports the average F1 score of different input combinations. Comparing the results on missing rates (0.8, 0.5, 0.5) and (0.8, 0.5, 0.2) (or (0.5, 0.5, 0.8) and (0.2, 0.5, 0.8)), it is clear that both MMIN and RedCore exhibit better performance with the former, even though the former provides less data. This confirms that more imbalanced missing rates (the latter) can lead to deteriorated performance, and MMIN is more sensitive to the imbalance. It is also shown that MMIN's lexical modality is sensitive to imbalanced modality missing. In comparison, RedCore showcases greater robustness against imbalanced missing rates. This is further supported by Figure 4 showing the average F1 score of all the baseline models on the three datasets.

Ablation Studies

In the ablation studies, we compare RedCore with two ablative settings on dataset IEMOCAP: Core — the Red part is excluded, namely, $\eta^1 = \eta^2 = \eta^3$; and Red — the Core part is partially removed, namely, VIB based individual modality label prediction is banned during training, and only the joint modality conducts prediction (the second term in eq. (2) only accounts for $m = M + 1$ instead of summing over
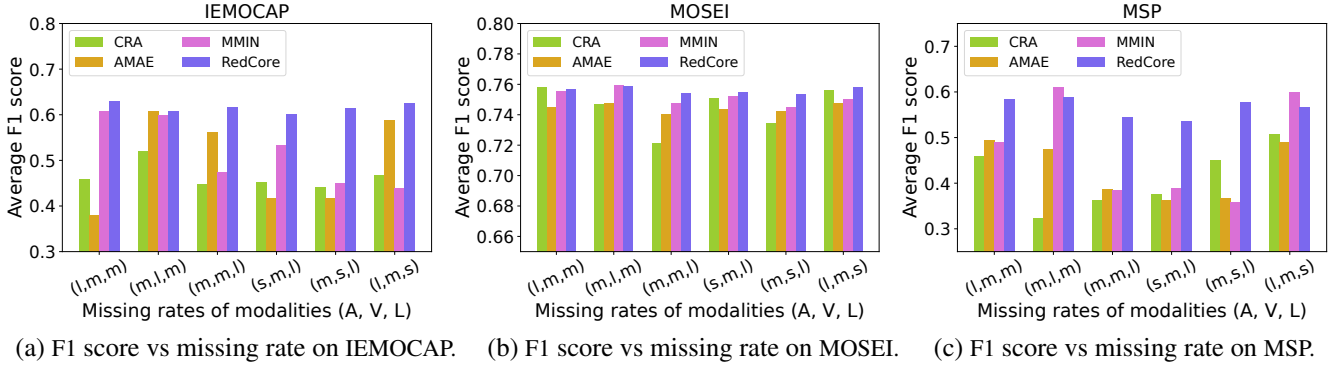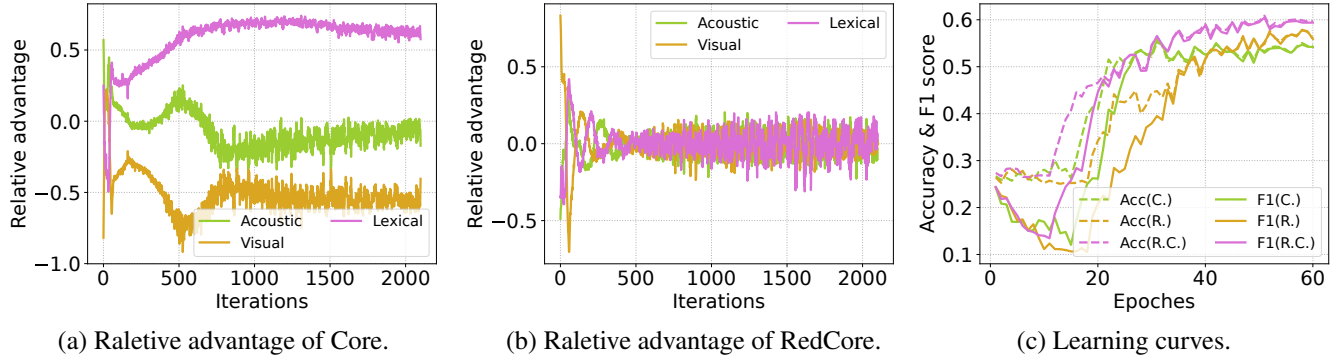
(a) F1 score vs missing rate on IEMOCAP.　(b) F1 score vs missing rate on MOSEI.　(c) F1 score vs missing rate on MSP.

Figure 4: Average F1 score with imbalanced missing rates for different datasets (l=0.8, m=0.5, and s=0.2).

(a) Raletive advantage of Core.　(b) Raletive advantage of RedCore.　(c) Learning curves.

Figure 5: Comparisons of Core, Red and RedCore on dataset IEMOCAP.

MR	(s,m,l)	(m,s,l)	(l,m,s)	(l,s,m)	(s,l,m)	(m,l,s)
C.	0.5696	0.5599	0.5653	0.5534	0.5481	0.5651
R.	0.5793	0.5708	0.5894	0.5831	0.5853	0.5849
R.C.	0.5999	0.6109	0.6017	0.5921	0.5993	0.5950

Table 2: Average F1 score with imbalanced missing rates (MR) for modalities (A,V,L) (l=0.8, m=0.5, and s=0.2).

m from 1 to $M + 1$). The models are trained with imbalanced missing rates and tested with all possible input modality combinations, and the average F1 score is reported in Table 2. Via comparing Core, Red and RedCore, it is obvious that VIB based cross-modal representation learning and relative advantage aware supervision regulation both contribute to the success of RedCore. (Due to the limited space in the tables and figures, we use C., R., and R.C. as the shorthand for Core, Red, and RedCore, respectively).

Figures 5(a) and 5(b) plot the relative advantages of the three modalities during training with missing rates (0.8, 0.2, 0.5) for Core and RedCore, respectively. Figure 5(a) shows that lexical modality dominates the other two when no supervision regulation is enforced. Comparing Figures 5(a) and 5(b), it is clear that supervision regulation achieves more balanced learning, indicating a smaller imputation loss gap between modalities. Shown in Figure 5(c) are the learning

curves (accuracy and F1 score) of the three settings, which demonstrates that RedCore is improved over Core and Red.

Conclusions

In this paper, we propose a cross-modal representation learning approach for missing modalities based on the VIB theory. Considering the imbalanced missing rates, a bi-level optimization problem is formulated to dynamically regulate the supervision of each modality during training. The numerical results validate the effectiveness of our design, and demonstrate the robustness of the proposed approach against modality missing. In our future work, we will investigate the convergence of algorithm 1 and analyze the Pareto optimality of the solution according to Lemma 1.

Acknowledgments

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

Bang, S.; Xie, P.; Lee, H.; Wu, W.; and Xing, E. 2021. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11396–11404.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; and Provost, E. M. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1): 67–80.

Chen, C.; Chen, X.; Ma, C.; Liu, Z.; and Liu, X. 2023. Gradient-based bi-level optimization for deep learning: A survey. *arXiv preprint arXiv:2207.11719v4*.

Du, C.; Du, C.; Wang, H.; Li, J.; Zheng, W.-L.; Lu, B.-L.; and He, H. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, 108–116.

Fernando, M.-P.; Cèsar, F.; David, N.; and José, H.-O. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7): 3217–3258.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

Hu, M.; and Chen, S. 2018. Doubly aligned incomplete multi-view clustering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2262–2268.

Jadhav, A.; Pramod, D.; and Ramanathan, K. 2019. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10): 913–933.

Jafrasteh, B.; Hernández-Lobato, D.; Lubián-López, S. P.; and Benavente-Fernández, I. 2023. Gaussian processes for missing value imputation. *Knowledge-Based Systems*, 273: 110603.

Jeong, H.; Wang, H.; and Calmon, F. P. 2022. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9558–9566.

Ji, K. 2021. *Bilevel optimization for machine learning: Algorithm design and convergence analysis*. The Ohio State University.

Josse, J.; Prost, N.; Scornet, E.; and Varoquaux, G. 2020. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Li, X.-L.; Chen, M.-S.; Wang, C.-D.; and Lai, J.-H. 2022. Refining graph structure for incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*.

Liang, J.; Li, R.; and Jin, Q. 2020. Semi-supervised multimodal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM international conference on multimedia*, 2852–2861.

Liu, H.; Wei, D.; Lu, D.; Sun, J.; Wang, L.; and Zheng, Y. 2023. M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. *arXiv preprint arXiv:2303.05302*.

Liu, J.; Teng, S.; Fei, L.; Zhang, W.; Fang, X.; Zhang, Z.; and Wu, N. 2021. A novel consensus learning approach to incomplete multi-view clustering. *Pattern Recognition*, 115: 107890.

Liu, X.; Li, M.; Tang, C.; Xia, J.; Xiong, J.; Liu, L.; Kloft, M.; and Zhu, E. 2020. Efficient and effective regularized incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2634–2646.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; and Gao, W. 2018. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(10): 2410–2423.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.

Manrique-Vallier, D.; and Reiter, J. P. 2017. Bayesian simultaneous edit and imputation for multivariate categorical data. *Journal of the American Statistical Association*, 112(520): 1708–1719.

Mao, H.; Yuan, Z.; Xu, H.; Yu, W.; Liu, Y.; and Gao, K. 2022. M-SENA: An Integrated Platform for Multimodal Sentiment Analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 204–213.

Miyaguchi, K.; Katsuki, T.; Koseki, A.; and Iwamori, T. 2022. Variational Inference for Discriminative Learning with Generative Modeling of Feature Incompletion. In *International Conference on Learning Representations*.

Morales-Alvarez, P.; Gong, W.; Lamb, A.; Woodhead, S.; Peyton Jones, S.; Pawlowski, N.; Allamanis, M.; and Zhang, C. 2022. Simultaneous missing value imputation and structure learning with groups. *Advances in Neural Information Processing Systems*, 35: 20011–20024.

Peng, X.; Wei, Y.; Deng, A.; Wang, D.; and Hu, D. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8238–8247.

Rezaei-Shoshtari, S.; Hogan, F. R.; Jenkin, M.; Meger, D.; and Dudek, G. 2021. Learning intuitive physics with multimodal generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6110–6118.

Sun, J.; Han, S.; Ruan, Y.-P.; Zhang, X.; Zheng, S.-K.; Liu, Y.; Huang, Y.; and Li, T. 2023. Layer-wise Fusion with Modality Independence Modeling for Multi-modal Emotion Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 658–670.

Sun, S.; Dong, W.; and Liu, Q. 2021. Multi-View Representation Learning With Deep Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 4453–4468.

Sun, Y.; Mai, S.; and Hu, H. 2021. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28: 1650–1654.

Tang, H.; and Liu, Y. 2022. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, 21090–21110. PMLR.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1405–1414.

Vu, T.; and Raich, R. 2022. On asymptotic linear convergence of projected gradient descent for constrained least squares. *IEEE Transactions on Signal Processing*, 70: 4061–4076.

Wang, Q.; Zhan, L.; Thompson, P.; and Zhou, J. 2020. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1828–1838.

Wang, W.; Tran, D.; and Feiszli, M. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12695–12705.

Wang, Y.; and Singh, L. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2): 101–119.

Wei, S.; Luo, C.; and Luo, Y. 2023. MMANet: Margin-aware Distillation and Modality-aware Regularization for Incomplete Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20039–20049.

Wen, J.; Zhang, Z.; Zhang, Z.; Zhu, L.; Fei, L.; Zhang, B.; and Xu, Y. 2021. Unified tensor framework for incomplete multi-view clustering and missing-view inferring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10273–10281.

Woo, S.; Lee, S.; Park, Y.; Nugroho, M. A.; and Kim, C. 2023. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2776–2784.

Wu, N.; Jastrzebski, S.; Cho, K.; and Geras, K. J. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, 24043–24055. PMLR.

Yang, J.; Ji, K.; and Liang, Y. 2021. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34: 13670–13682.

Yoon, S.; and Sull, S. 2020. GAMIN: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8456–8464.

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

Zhang, C.; Li, H.; Chen, C.; Jia, X.; and Chen, C. 2022. Low-Rank Tensor Regularized Views Recovery for Incomplete Multiview Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Zhang, Y.; and Long, Q. 2021. Assessing fairness in the presence of missing data. *Advances in neural information processing systems*, 34: 16007–16019.

Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2608–2618.