Boomda: Balanced Multi-objective Optimization for Multimodal Domain Adaptation

Jun Sun[1], **Xinxin Zhang**[2], **Simin Hong, Jian Zhu**[1*], **Xiang Gao**[1*]

[1]Zhejinag Lab, Hangzhou, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Email: sunjun16sj@gmail.com [*]Corresponding author

Abstract

Multimodal learning, while contributing to numerous success stories across various fields, faces the challenge of prohibitively expensive manual annotation. To address the scarcity of annotated data, a popular solution is unsupervised domain adaptation, which has been extensively studied in unimodal settings yet remains less explored in multimodal settings. In this paper, we investigate heterogeneous multimodal domain adaptation, where the primary challenge is the varying domain shifts of different modalities from the source to the target domain. We first introduce the information bottleneck method to learn representations for each modality independently, and then match the source and target domains in the representation space with correlation alignment. To balance the domain alignment of all modalities, we formulate the problem as a multi-objective task, aiming for a Pareto optimal solution. By exploiting the properties specific to our model, the problem can be simplified to a quadratic programming problem. Further approximation yields a closed-form solution, leading to an efficient modality-balanced multimodal domain adaptation algorithm. The proposed method features **B**alanced multi-**o**bjective **o**ptimization for **m**ultimodal **d**omain **a**daptation, termed **Boomda**. Extensive empirical results showcase the effectiveness of the proposed approach and demonstrate that Boomda outperforms the competing schemes. The code is is available at: https://github.com/sunjunaimer/Boomda.git.

1 Introduction

Multimodal learning typically leverages heterogeneous and complementary signals, such as acoustic, visual, and linguistic information, to perform machine learning tasks like classification, clustering and retrieval. Thanks to recent advances in hardware and model design, multimodal learning has been applied to various applications, including but not limited to action recognition (Woo et al. 2023), affective computing (Sun et al. 2023; Zhang et al. 2024), and medical analysis (Liu et al. 2023). Compared with the unimodal alternative, multimodal learning achieves remarkable performance improvement. However, one of its notorious drawbacks is that collecting and annotating multimodal data is expensive and time-consuming. Consequently, the scarcity of annotated data presents a major challenge for the practical application of multimodal learning.

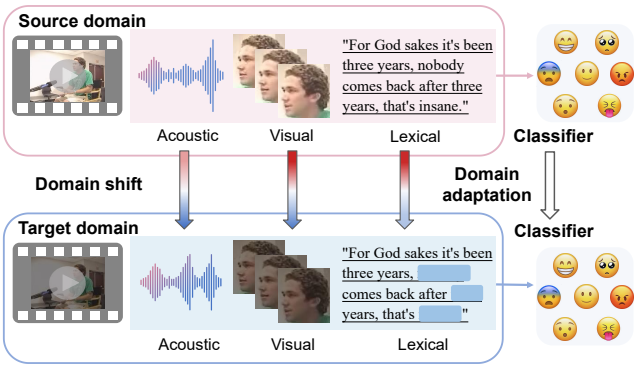


Figure 1: Domain shift in the context of multimodal emotion recognition. The example sample is drawn from dataset IEMOCAP (Busso et al. 2008).

To alleviate this issue, unsupervised domain adaptation has merged as a powerful approach for enhancing the model's ability to transfer knowledge from a label-rich source domain to an unlabeled but related target domain. The objective of unsupervised domain adaptation is to train a model on the source domain while optimizing its performance on the target domain, which lacks labels during training. Mainstream approaches usually align source and target domains and learn generalizable representations via either directly minimizing the feature distribution shift between the two domains, or training an adversarial domain classifier. A massive amount of techniques following these principles have been developed, demonstrating impressive results across various tasks, including image classification (Hu et al. 2023; Lee et al. 2023), semantic segmentation (Wang et al. 2023), object detection (Kennerley et al. 2023), question answering (Dua et al. 2023), among others.

Nevertheless, the existing domain adaptation literature primarily focuses on unimodal settings, particularly within the areas of computer vision and natural language processing. In comparison, multimodal domain adaptation remains less investigated, yet is attracting growing interests due to the popularity of multimodal learning (Sun et al. 2025). From the perspective of modality diversity, multimodal domain adaptation is conceptually divided into two cate-

gories: homogeneous and heterogeneous multimodal adaptation (Singhal et al. 2023). The former focuses on scenarios where multiple modalities share similar underlying structure or environment. For instance, in multimodal visual domain adaptation tasks, different modalities refer to 2D image and 3D point cloud (Xing et al. 2023; Liu et al. 2021), optical flow and RGB image (Munro and Damen 2020),or CT and MRI image (Kruse, Hansen, and Heinrich 2021). In such cases, the gaps between source and target domains of different modalities are relatively small, suggesting that the domains of different modalities can be aligned uniformly.

The latter, heterogeneous multimodal domain adaptation, deals with scenarios where different modalities exhibit diverse forms and reside in separate spaces. A typical example is the multimodal emotion recognition task as illustrated in Figure 1, where acoustic, visual and lexical modalities are utilized simultaneously to detect emotions. Each modality faces distinct domain shift factors (e.g., background noise, illumination, and talking scenarios for the three modalities, respectively), naturally leading to varying degrees of distribution shifts from the source to the target domain across different modalities.

In this paper, we investigate heterogeneous multimodal domain adaptation, which is more challenging than its homogeneous counterpart, as analyzed above. Directly applying current unimodal domain adaptation techniques to the heterogeneous multimodal setting potentially causes an imbalanced alignment of different modalities. The imbalance can result in some modalities dominating the training process and being well-trained, while others remaining undertrained (Fan et al. 2023; Sun et al. 2024). Henceforth, it is critical to balance the training for multimodal models in order to fully take advantage of all modalities.

Towards the goal of achieving balanced multimodal domain adaptation, this work develops a model framework building upon information bottleneck (IB) theory (Saxe et al. 2019; Kawaguchi et al. 2023) to align the representations of the source and target domains, and proposes an approach to balance the alignment across modalities using the multi-objective optimization scheme. Specifically, we first construct the model with pretrained backbones for each modality. In order to retain general knowledge while adapting to new tasks and domains, the pretrained backbones are partially finetuned with some layers frozen. Then, we apply IB theory to formulate the training objective, promoting modality independence. The IB-based method ensures that each modality performs label prediction, thereby encouraging each to develop its optimal representation independently. Subsequently, the representations of each modality between the source and target domains are matched using correlation alignment (Coral) (Sun, Feng, and Saenko 2016).

The objectives of aligning all modalities are competing with each other, which hence can be formulated as a multi-objective optimization (Navon et al. 2022; Hu et al. 2024). To balance the alignment losses among all modalities, we employ multiple gradient descent algorithm (MGDA) (Sener and Koltun 2018) to obtain the Pareto optimal solution. This algorithm involves solving an optimization problem at each training iteration. Our analysis reveals that for our model structure, this optimization problem enjoys a desirable property, which facilitates the formulation of an approximated problem. Fortunately, the approximated problem admits a closed-form solution, and thus results in an efficient training algorithm.

In summary, the present paper proposes a novel approach — **B**alanced multi-**o**bjective **o**ptimization for **m**ultimodal **d**omain **a**daptation, dubbed **Boomda**. The primary contributions are as follows:

1. We develop a multimodal domain adaptation framework, where each modality is encouraged to learn its optimal representation separately based on the IB theory. Then, the source and target domains are aligned in the representation space with correlation alignment.
2. To balance all modalities, we formulate the alignment for them as a multi-objective optimization problem, which can be solved efficiently using MGDA algorithm via exploiting a favorable property of our problem.
3. Extensive experiments conducted on widely used benchmark datasets demonstrate the superior performance of Boomda compared to previous schemes.

2 Related Works

Domain adaptation

A plethora of prior studies have been devoted to domain adaptation, and survey papers (Wang and Deng 2018; Li et al. 2024a) offer a comprehensive review of them. In this section, we cover the most relevant literature, which can be broadly classified into two groups: adversarial-based methods and moment-based methods. Adversarial-based methods center on learning domain-invariant features by an adversarial scheme where a domain discriminator is trained against the feature extractor. The earliest work that demonstrates the effectiveness of adversarial learning for domain adaptation is DANN (Ganin et al. 2016), following which numerous adversarial-based methods emerge. MDAN (Zhao et al. 2018) investigates multiple source domain adaptation and devises two versions of optimization strategies. Label prediction information is introduced as conditioning for domain alignment in CDAN (Long et al. 2018) and MADA (Pei et al. 2018). CDA (Yadav et al. 2023) integrates contrastive learning into domain adaptation to achieve class-level alignment. DADA (Tang and Jia 2020) combines domain and category classifiers as a shared classifier to encourage a mutually inhibitory relation between domain and category predictions. Similarly, DALN (Chen et al. 2022a) develops a discriminator-free adversarial model via reusing the task-specific classifier as a discriminator. Other adversarial methods, PCL (Li et al. 2024b) and DADA(Ren et al. 2024), incorporate data augmentation from the raw feature space and representation space, respectively.

Moment-based methods mitigate domain shifts by minimizing moment-based distribution discrepancy across domains. Maximum mean discrepancy (MMD) based methods, such as DDC (Tzeng et al. 2014) and MK-MMD (Long et al. 2015), represent first-order moment approaches which match the mean of the representations. DDC directly minimizes the maximum mean discrepancy of the source and

target domains, and MK-MMD incorporates a multi-kernel strategy to enhance DDC. Coral (Sun, Feng, and Saenko 2016; Sun and Saenko 2016) and JDDA (Chen et al. 2019) are typical second-order moment approaches, matching the covariance of the representations. Coral aligns the correlation matrix of features from two domains. Built on Coral, JDDA aims to attain class-distinct features via additionally developing instance-based and center-based discriminative learning strategies.

Imbalance in multimodal learning

The imbalance issue is naturally encountered in multimodal learning and is receiving increasing research attention. In studies (Wang, Tran, and Feiszli 2020) and (Peng et al. 2022), it is observed that different modalities can generalize and overfit at different rates, and some strong modalities might dominate the training and suppress other weak modalities. This inspires works (Wang, Tran, and Feiszli 2020), (Wu et al. 2022) and (Sun, Mai, and Hu 2021; Peng et al. 2022; Sun et al. 2024) to dynamically regulate the learning rates of different modalities during training according to the overfitting estimation, the gradient, and the loss advantage of each modality, respectively. Besides this line, knowledge distillation is employed in work (Du et al. 2021) to reinforce the unimodal modules from pretrained teacher models and thus prevent modality failure. Self-distillation is utilized to balance the optimization objectives of all modalities in work (Shi et al. 2024).

In our work, we adopt the second-order moment matching method for representation alignment, thus focusing on balancing modalities and circumventing the difficulty in balancing the competing generative and discriminative components in adversarial-based methods. We introduce a multi-objective optimization method to balance the alignment of all modalities.

3 Method: Boomda

Before introducing our method, Boomda, we first define the notations to be used.

Notations: Assume in a multimodal classification task, the training dataset consists of N samples, each with M modalities. For simplicity, we introduce an auxiliary modality containing all modalities, resulting in a total of $M+1$ modalities. Let $[I]$ for any positive integer I denote the set $\{1, 2, \ldots, I\}$. The training samples are represented by $(\{\boldsymbol{x}_{n,m}\}_{m\in[M]}, \{\boldsymbol{y}_{n,m}\}_{m\in[M+1]})$, where $n \in [N]$ is the index of the samples, $\boldsymbol{x}_{n,m} \in \mathbb{R}^{d_m}$ represents the d_m-dimensional feature vector (or vector sequence) of modality m, $\forall m \in [M]$, and $\boldsymbol{y}_{n,m}$ denotes the label associated with modality m, $\forall m \in [M+1]$. In datasets where all modalities share a common label, $\boldsymbol{y}_{n,1} = \boldsymbol{y}_{n,2} = \cdots = \boldsymbol{y}_{n,M+1}$ holds. Suppose there are C categories in the classification task; then label $\boldsymbol{y}_{n,m}$ can be either a one-hot vector or a scalar in $[C]$, and we adopt either form as necessary in the rest of the paper. For consistency, we use $\boldsymbol{x}_{n,M+1} := [\boldsymbol{x}_{n,1}; \boldsymbol{x}_{n,2}; \cdots; \boldsymbol{x}_{n,M}]$ to aggregate all modalities.

Let $\boldsymbol{X}_m$ and $\boldsymbol{Y}_m$ be general feature and label random variables for all $m \in [M+1]$, with $\boldsymbol{x}_{n,m}$ and $\boldsymbol{y}_{n,m}$ being their specific instances. Let $\boldsymbol{Z}_m \in \mathbb{R}^d$, derived from $\boldsymbol{X}_m$, denote the representation of modality m; and $\boldsymbol{z}_{n,m}$ is an instance of $\boldsymbol{Z}_m$ (for brevity, we assume the representations of all modalities are d-dimensional vectors). Superscripts s and t are used to differentiate variables from the source and target domains. For example, $\boldsymbol{X}_m^s$ and $\boldsymbol{X}_m^t$ represent the features of modality m from the two domains, respectively.

In the sequel, we will first present our model design and derive the training objective function, including the model framework, IB based representation learning, pseudo labeling approach, and modality-wise representation alignment. Then, we will balance the alignment across modalities via casting it as a multi-objective optimization problem, for which we develop an efficient algorithm.

Multimodal Domain Adaptation

A. Model Framework Overview. In this section, we focus on the model framework and ignore the implementation details, which will be elaborated later in the Numerical Results section. Figure 2 illustrates the proposed multimodal domain adaptation framework in an example with two modalities (visual and acoustic), namely, $M = 2$. The raw features $\boldsymbol{X}_m, \forall m \in [M]$ are first tokenized and fed into the pretrained transformer-based models, of which the top layers will be finetuned. Following the pretrained models are sequence encoders which further encode the sequence features into vector representations $\boldsymbol{Z}_m, \forall m \in [M]$. More formally, for each modality $m \in [M]$, the corresponding pretrained model and the sequence encoder can be summarized by a deterministic encoder function $f_m^e(\cdot; \boldsymbol{\theta}_m^e) : \mathbb{R}^{d_m} \to \mathbb{R}^d$ with trainable parameter $\boldsymbol{\theta}_m^e$; then we have $\boldsymbol{Z}_M = f_m^e(\boldsymbol{X}_m; \boldsymbol{\theta}_m^e)$. The multimodal representation is denoted by $\boldsymbol{Z}_{M+1} := [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_M]$, a concatenation of the representations of all modalities.

Each modality $m \in [M+1]$ is associated with a classifier $f_m^c(\cdot, \boldsymbol{\theta}_m^c)$ with parameter $\boldsymbol{\theta}_m^c$ for label prediction; that is, $\hat{\boldsymbol{Y}}_m = f_m^c(\boldsymbol{Z}_m, \boldsymbol{\theta}_m^c)$. The multimodal prediction $\hat{\boldsymbol{Y}}_{M+1}$ is assigned to be the ultimate predicted label. For brevity of expression, we use $\boldsymbol{\theta} := \{\boldsymbol{\theta}_{M+1}^c\} \cup \{\boldsymbol{\theta}_m^e, \theta_m^c\}_{m\in[M]}$ to collect all model parameters, and $\boldsymbol{\theta}^e := \{\boldsymbol{\theta}_m^e\}_{m\in[M]}$ to collect the parameters of all encoders.

B. IB based Representation Learning. With the above model framework, the information follows $\boldsymbol{X}_m \to \boldsymbol{Z}_m \to \boldsymbol{Y}_m, \forall m \in [M+1]$. Information bottleneck based representation learning aims to obtain representations $\boldsymbol{Z}_m^s, \forall m \in [M+1]$, such that they captures generalizable features, and thereby alleviates the difficulty in aligning the source and target representations.

Mathematically, the optimal representation is generated by minimizing the following IB loss on the source domain:

$$\mathcal{L}^{IB}(\boldsymbol{\theta}) := \sum_{m\in[M+1]} \beta I(\boldsymbol{X}_m^s, \boldsymbol{Z}_m^s) - I(\boldsymbol{Z}_m^s, \boldsymbol{Y}_m^s), \quad (1)$$

where $I(\cdot, \cdot)$ denotes the mutual information of any two random variables, and β is a predefined coefficient.

From the perspective of information theory, it is obvious that the resultant representation $\boldsymbol{Z}_m^s$ retains minimal information from the raw feature $\boldsymbol{X}_m^s$ yet maintains the maximal

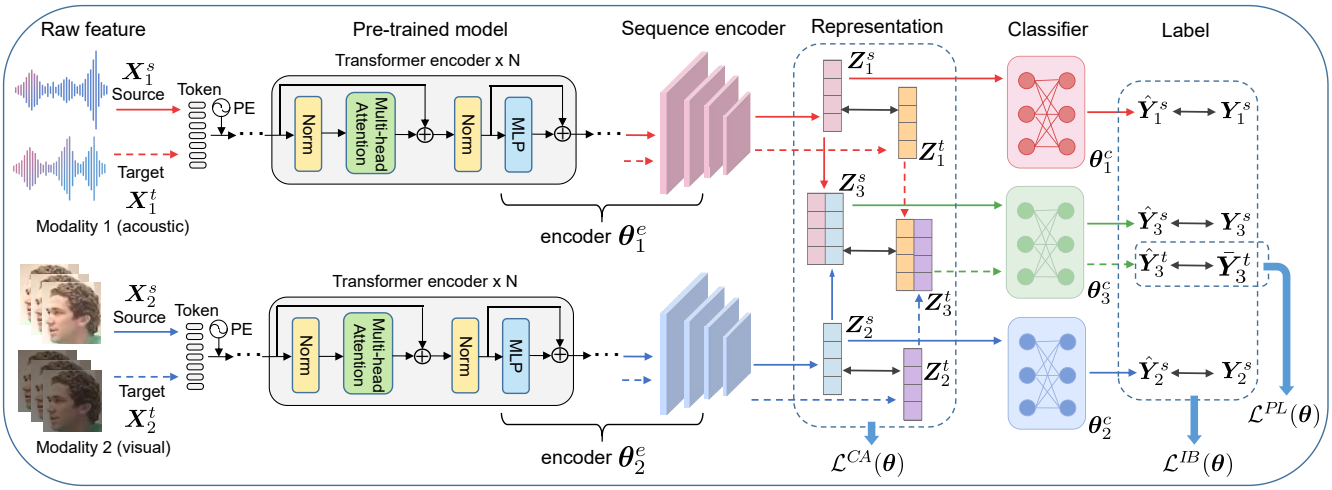


Figure 2: Model framework with 2 modalities as an example (multimodal representation $\boldsymbol{Z}_3$ is a concatenation of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$; solid and dashed regular arrows represent the flows of source and target domains, respectively; double-headed arrows represent alignment or supervision signals, corresponding to the information bottleneck loss $\mathcal{L}^{IB}(\boldsymbol{\theta})$, pseudo label supervision loss $\mathcal{L}^{PL}(\boldsymbol{\theta})$ and correlation alignment loss $\mathcal{L}^{CA}(\boldsymbol{\theta})$).

information of the label $\boldsymbol{Y}_m^s$. Therefore, $\boldsymbol{Z}_m^s$ is an optimal representation in the sense of information bottleneck theory (Saxe et al. 2019; Kawaguchi et al. 2023). Moreover, each individual modality m, for all $m \in [M]$, is enforced to generate its own optimal representation, which promotes modality independence and prevents some weak modalities from being dominated by strong ones.

Then, we specify how the two information terms in Eq. (1) are computed.

$$\begin{aligned} I(\boldsymbol{X}_m^s, \boldsymbol{Z}_m^s) &= H(\boldsymbol{Z}_m^s) - H(\boldsymbol{Z}_m^s|\boldsymbol{X}_m^s) \\ &= H(\boldsymbol{Z}_m^s) = \mathbb{E}_{\boldsymbol{Z}_m^s}[-\log p(\boldsymbol{Z}_m^s)], \end{aligned} \tag{2}$$

where $H(\cdot)$ represents entropy, and $H(\boldsymbol{Z}_m^s|\boldsymbol{X}_m^s) = 0$ since $\boldsymbol{Z}_m = f_m^e(\boldsymbol{X}_m; \boldsymbol{\theta}_m^e)$ is a deterministic function. It is a convention to assume that $p(\boldsymbol{Z}_m^s)$ follows Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_m^s, \boldsymbol{\Sigma}_m^s)$ ($\boldsymbol{\mu}_m^s \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}_m^s \in \mathbb{R}^{d\times d}$ is a diagonal matrix). As a result, we can estimate $\boldsymbol{\mu}_m^s$ and $\boldsymbol{\Sigma}_m^s$ with the representations $\boldsymbol{z}_{n,m}^s, n \in [N^s]$, and thus the entropy of $H(\boldsymbol{Z}_m^s)$ can be obtained as:

$$H(\boldsymbol{Z}_m^s) = \frac{1}{2}\log|\boldsymbol{\Sigma}_m^s| + \frac{d}{2}(1 + \log(2\pi)), \tag{3}$$

where $|\boldsymbol{\Sigma}_m^s|$ represents the determinant of $\boldsymbol{\Sigma}_m^s$.

Similarly, $I(\boldsymbol{Z}_m^s, \boldsymbol{Y}_m^s)$ can be written as:

$$\begin{aligned} I(\boldsymbol{Z}_m^s, Y_m^s) &= H(\boldsymbol{Y}_m^s) - H(\boldsymbol{Y}_m^s|\boldsymbol{Z}_m^s) \\ &= H_{Y,m}^s - H(\boldsymbol{Y}_m^s|\boldsymbol{Z}_m^s) \\ &= H_{Y,m}^s + \frac{1}{N^s}\sum_{n=1}^{N^s}\log p(\boldsymbol{y}_{n,m}^s|\boldsymbol{z}_{n,m}^s), \end{aligned} \tag{4}$$

where $H(\boldsymbol{Y}_m^s) = H_{Y,m}^s$ is a constant independent from the model parameter $\boldsymbol{\theta}$.

Combining Eqs. (1), (2), (3) and (4) gives the information bottleneck loss as follows (with constant terms omitted):

$$\begin{aligned} \mathcal{L}^{IB}(\boldsymbol{\theta}) = \sum_{m=1}^{M+1}\Big[\frac{\beta}{2}\log|\boldsymbol{\Sigma}_m^s| \\ - \frac{1}{N^s}\sum_{n\in[N^s]}\log p(\boldsymbol{y}_{n,m}^s|\boldsymbol{z}_{n,m}^s)\Big], \end{aligned} \tag{5}$$

where the first term is a regularization for the representation that suppresses the noisy and ineffective information; and the second term corresponds to the negative log-likelihood of the prediction (equivalent to cross-entropy loss).

C. Pseudo Labeling by Voting. We introduce a voting strategy to yield pseudo labels for the target domain samples to supervise label prediction. Given the predictions $\hat{\boldsymbol{y}}_{n,m}^t$ of all modalities $m \in [M+1]$ (here $\hat{\boldsymbol{y}}_{n,m}^t$ is supposed to be one-hot vector), the voting result is:

$$\hat{\boldsymbol{y}}_n^t = \sum_{m\in[M+1]}\hat{\boldsymbol{y}}_{n,m}^t. \tag{6}$$

The c-th element of $\hat{\boldsymbol{y}}_n^t$ is $(\hat{\boldsymbol{y}}_n^t)_c, \forall c \in [C]$, denoting the number of votes for sample n being classified as category c.

Subsequently, to ensure the reliability of labeling, only the samples that receive at least M_v votes for the same category are selected to constitute the pseudo labeling set $\mathcal{N}_v^t$:

$$\mathcal{N}_v^t = \{n|\max\{(\hat{\boldsymbol{y}}_n^t)_1, \cdots, (\hat{\boldsymbol{y}}_n^t)_C\} \geq M_v, n \in [N^t]\}. \tag{7}$$

The pseudo labels for the selected samples are attained as:

$$\bar{\boldsymbol{y}}_n^t = \text{argmax}_c\{(\hat{\boldsymbol{y}}_n^t)_c|c = 1, 2, \cdots, C\}, \forall n \in \mathcal{N}_v^t, \tag{8}$$

where $\bar{\boldsymbol{y}}_n^t$ can also adopt the corresponding one-hot vector form when necessary.

Then the pseudo labels serve as supervision signals for training on the target domain, meaning the following cross-entropy loss is minimized:

$$\mathcal{L}^{PL}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{N}_v^t|}\sum_{n\in\mathcal{N}_v^t}\sum_{c\in[C]} -(\bar{\boldsymbol{y}}_n^t)_c \log(\hat{\boldsymbol{y}}_{n,M+1}^t)_c, \tag{9}$$

where $|\cdot|$ represents the cardinality of a set, and $(\cdot)_c$ is the c-th element of a vector.

D. Representation Alignment. We align the source and target domains for each modality in its own representation space separately. Specifically, we first calculate the correlation matrices of $\boldsymbol{Z}_m^s$ and $\boldsymbol{Z}_m^t$, denoting them as $\boldsymbol{C}_m^s$ and $\boldsymbol{C}_m^t$, respectively, for $m \in [M+1]$. Then, we match the representations by minimizing the following correlation alignment (CA) loss (Sun, Feng, and Saenko 2016):

$$\mathcal{L}_m^{CA}(\boldsymbol{\theta}) = ||\boldsymbol{C}_m^t - \boldsymbol{C}_m^s||_F^2, \tag{10}$$

where $||\cdot||_F$ means the Frobenius norm. Let $\boldsymbol{L}^{CA}(\boldsymbol{\theta}) := [\mathcal{L}_1^{CA}(\boldsymbol{\theta}), \mathcal{L}_2^{CA}(\boldsymbol{\theta}), \cdots, \mathcal{L}_{M+1}^{CA}(\boldsymbol{\theta})]^T$ collect the representation alignment losses for all modalities. Then, the overall loss for model training writes:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^{IB}(\boldsymbol{\theta}) + \alpha_1 \mathcal{L}^{PL}(\boldsymbol{\theta}) + \alpha_2 h(\boldsymbol{L}^{CA}(\boldsymbol{\theta})), \tag{11}$$

where α_1 and α_2 are two constant coefficients balancing the losses. Note that without incurring confusion, the formula in Eq. (11) is not mathematically rigorous, since $h(\boldsymbol{L}^{CA}(\boldsymbol{\theta}))$ can be a vector-valued function as will be shown later.

A straightforward method to aggregate the alignment losses of all modalities is to define $h(\boldsymbol{L}^{CA}(\boldsymbol{\theta}))$ as a weighted sum of the losses. However, the downsides come in order: 1) without prior expert knowledge, it is difficult to specify the weights for all modalities; 2) searching for optimal weights becomes expensive when dealing with a large number of modalities; 3) predetermined fixed weights might not adapt well to the dynamic nature of alignment losses during training. To address these limitations, in the sequel, we will cast the multimodal alignment problem as a multi-objective optimization problem and develop an efficient algorithm to solve it.

Pareto optimal balance across modalities

A. Multi-objective Optimization for Multimodal Domain Adaptation. The multimodal domain adaptation problem can be formulated as a multi-objective optimization problem, where multiple potentially competing objectives are optimized simultaneously. Concretely, the problem is formalized as minimizing a vector-valued loss:

$$\min_{\boldsymbol{\theta}} h(\boldsymbol{L}^{CA}(\boldsymbol{\theta})) := [\mathcal{L}_1^{CA}(\boldsymbol{\theta}), \mathcal{L}_2^{CA}(\boldsymbol{\theta}), \cdots, \mathcal{L}_{M+1}^{CA}(\boldsymbol{\theta})]^T. \tag{12}$$

The goal of multi-objective optimization centers on finding a Pareto optimal solution, as defined below.

Definition 1. *(Pareto Optimality) (a) A solution $\boldsymbol{\theta}$ dominates another solution $\boldsymbol{\theta}'$ if $\boldsymbol{L}^{CA}(\boldsymbol{\theta}) \neq \boldsymbol{L}^{CA}(\boldsymbol{\theta}')$ and $\mathcal{L}_m^{CA}(\boldsymbol{\theta}) \leq \mathcal{L}_m^{CA}(\boldsymbol{\theta}')$ hold for all $m \in \{1, 2, \cdots, M\}$; (b) A solution $\boldsymbol{\theta}^*$ is Pareto optimal if there exists no solution that dominates $\boldsymbol{\theta}^*$.*

As per works (Désidéri 2012) and (Sener and Koltun 2018), we adopt the multiple gradient descent algorithm (MGDA) to solve problem (12). MGDA is built upon the Karush-Kuhn-Tucker (KKT) conditions, which are necessary conditions for the solution to be Pareto optimal and defines the Pareto stationary point. Specifically, any solution $\boldsymbol{\theta}$ is called a Pareto stationary point if there exists a vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_{M+1}]^T$, such that the following conditions are satisfied: a) $\boldsymbol{\gamma} \geq \mathbf{0}$, b) $\mathbf{1}^T \cdot \boldsymbol{\gamma} = 1$, and c) $\sum_{m\in[M+1]} \gamma_m \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{1}$ and $\mathbf{0}$ represent all-one and all-zero vector of proper size, respectively.

Then, to find Pareto stationary point involves solving the problem below:

$$\mathbf{P1}: \min_{\boldsymbol{\gamma}} \; || \sum_{m\in[M+1]} \gamma_m \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta})||_2^2$$
$$s.t. \;\; \boldsymbol{\gamma} \geq \mathbf{0}, \;\; \mathbf{1}^T \cdot \boldsymbol{\gamma} = 1.$$

Suppose that $\boldsymbol{\gamma}^*$ is the optimal solution of problem **P1**, then two outcomes follow: 1) $\sum_{m\in[M+1]} \gamma_m^* \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) = \mathbf{0}$, which means that the corresponding $\boldsymbol{\theta}$ is a Pareto stationary point, and 2) $\sum_{m\in[M+1]} \gamma_m^* \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) \neq \mathbf{0}$, which is a descent direction for all objectives, and thus can be applied to update the model parameter. Equivalently, objective $\sum_{m\in[M+1]} \gamma_m^* \mathcal{L}_m^{CA}(\boldsymbol{\theta})$ is optimized, where $\boldsymbol{\gamma}^*$ is the coefficient that balances all modalities and leads to a Pareto stationary point.

Following the work of (Sener and Koltun 2018), we have a more computationally efficient version of problem **P1**:

$$\mathbf{P2}: \min_{\boldsymbol{\gamma}} \; || \sum_{m\in[M+1]} \gamma_m \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta})||_2^2$$
$$\text{s.t.} \;\; \boldsymbol{\gamma} \geq \mathbf{0}, \;\; \mathbf{1}^T \cdot \boldsymbol{\gamma} = 1.$$

Due to space limitations, we delegate the detailed derivation of problem **P2** to the appendix in the supplementary material. In contrast to problem **P1**, problem **P2** only requires the computation of $\nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta})$ instead of $\nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta})$, which reduces computational cost significantly, especially for deep neural networks.

B. Efficient Balanced Multimodal Domain Adaptation. In this section, we further explore the properties of problem **P2** specific to our model, which will facilitate the design of an efficient modality-balanced algorithm. For brevity, we define $\boldsymbol{g}_m := \nabla_{\boldsymbol{Z}_m} \mathcal{L}_m^{CA}(\boldsymbol{\theta})$ and $\boldsymbol{g}_{M+1}^m := \nabla_{\boldsymbol{Z}_m} \mathcal{L}_{M+1}^{CA}(\boldsymbol{\theta}), \forall m \in [M]$. Then, it can be verified that for the proposed model, the gradients take the following forms:

$$\boldsymbol{P} := \begin{bmatrix} \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_1^{CA}(\boldsymbol{\theta}) \\ \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_2^{CA}(\boldsymbol{\theta}) \\ \vdots \\ \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_{M+1}^{CA}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{g}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{g}_{M+1}^1 & \boldsymbol{g}_{M+1}^2 & \cdots & \boldsymbol{g}_{M+1}^M \end{bmatrix}.$$

With the above notations, problem **P2** can be equivalently written in a more compact manner:

$$\mathbf{P3}: \min_{\boldsymbol{\gamma}} \;\; \boldsymbol{\gamma}^T \boldsymbol{P}\boldsymbol{P}^T \boldsymbol{\gamma} = \boldsymbol{\gamma}^T \boldsymbol{Q} \boldsymbol{\gamma}$$
$$\text{s.t.} \;\; \boldsymbol{\gamma} \geq \mathbf{0}, \;\; \mathbf{1}^T \cdot \boldsymbol{\gamma} = 1,$$

where $\boldsymbol{Q} := \boldsymbol{P}\boldsymbol{P}^T =$

$$\begin{bmatrix} \boldsymbol{g}_1^T \boldsymbol{g}_1 & 0 & \cdots & 0 & \boldsymbol{g}_1^T \boldsymbol{g}_{M+1}^1 \\ 0 & \boldsymbol{g}_2^T \boldsymbol{g}_2 & \cdots & 0 & \boldsymbol{g}_2^T \boldsymbol{g}_{M+1}^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{g}_M^T \boldsymbol{g}_M & \boldsymbol{g}_M^T \boldsymbol{g}_{M+1}^M \\ \boldsymbol{g}_1^T \boldsymbol{g}_{M+1}^1 & \boldsymbol{g}_2^T \boldsymbol{g}_{M+1}^2 & \cdots & \boldsymbol{g}_M^T \boldsymbol{g}_{M+1}^M & \sum_{m=1}^M (\boldsymbol{g}_{M+1}^m)^T \boldsymbol{g}_{M+1}^m \end{bmatrix}.$$

Methods	IEMOCAP					MSP-IMPROV				
	AL	AV	VL	AVL	ave.	AL	AV	VL	AVL	ave.
D.T.	30.99	39.34	36.37	40.31	36.75	25.09	34.48	35.77	38.76	33.53
DANN	37.32	43.96	39.87	45.22	41.59	32.69	**42.44**	37.72	37.41	37.57
DADA	41.87	40.27	35.63	45.16	40.73	30.87	39.59	34.97	42.94	37.09
MADA	38.17	35.35	42.32	49.85	41.42	29.86	41.32	43.38	44.13	39.67
DALN	40.41	47.23	**42.99**	45.58	44.05	30.36	35.68	42.23	40.29	37.14
PCL	44.85	47.01	41.44	49.72	45.76	32.27	41.63	41.19	45.02	40.02
CDAN	47.23	47.36	41.74	51.47	46.95	32.33	40.56	42.30	45.21	40.07
Boomda	**49.81**	**47.46**	42.83	**54.82**	**48.73**	**33.30**	40.31	**45.10**	**47.29**	**41.50**

Table 1: The performance comparisons (in terms of F1 score) of Boomda and the existing methods.

Algorithm 1: Balanced multimodal domain adaptation

Initialization: initialize model parameter $\boldsymbol{\theta}^0$.
for $k = 0$ to $K-1$ **do**
  (1) perform forward pass of the model, and generate the pseudo labels according to Eqs. (7) and (8);
  (2) perform two backward passes of the model to obtain matrix $\boldsymbol{Q}$ and its diagonal approximation $\tilde{\boldsymbol{Q}}$;
  (3) solve quadratic problem **P3** or use the close-form solution Eq. (13) to attain the weight $\boldsymbol{\gamma}^k$;
  (4) calculate the overall loss $\mathcal{L}(\boldsymbol{\theta}^k)$ as Eq. (14);
  update the model parameter using an optimizer (e.g., Adam): $\boldsymbol{\theta}^{k+1} \leftarrow \text{optimizer}(\boldsymbol{\theta}^k, \eta)$.
end for
Return: Model parameter $\boldsymbol{\theta}^K$.

In experiments, it is observed that the absolute values of the off-diagonal entries of $\boldsymbol{Q}$ are much smaller than the diagonal entries, indicating that matrix $\boldsymbol{Q}$ is positive definite. The immediate result is that problem **P3** is a standard quadratic programming problem and can be solved using an off-the-shelf solver.

Furthermore, we can approximate $\boldsymbol{Q}$ with its diagonal matrix $\tilde{\boldsymbol{Q}}$, and solve the following problem:

$$\textbf{P4}: \quad \min_{\boldsymbol{\gamma}} \quad \boldsymbol{\gamma}^T \tilde{\boldsymbol{Q}} \boldsymbol{\gamma} \\ \text{s.t.} \quad \boldsymbol{\gamma} \geq \mathbf{0}, \quad \mathbf{1}^T \cdot \boldsymbol{\gamma} = 1.$$

A desirable property of problem **P4** follows.

Theorem 1. *Problem **P4** admits a closed-form solution:*

$$\boldsymbol{\gamma} = \frac{\tilde{\boldsymbol{Q}}^{-1}\mathbf{1}}{\mathbf{1}^T \tilde{\boldsymbol{Q}}^{-1}\mathbf{1}}. \tag{13}$$

We provide the proof of **Theorem 1** for completeness in the appendix. Upon obtaining coefficient $\boldsymbol{\gamma}$, the overall loss in Eq. (11) boils down to:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^{IB}(\boldsymbol{\theta}) + \alpha_1 \mathcal{L}^{PL}(\boldsymbol{\theta}) + \alpha_2 \sum_{m \in [M+1]} \gamma_m \mathcal{L}_m^{CA}(\boldsymbol{\theta}). \tag{14}$$

To sum up, the proposed modality-balanced multimodal domain adaptation approach is outlined in **Algorithm 1**. At each iteration, the pseudo labeling set and the associated labels are first generated. Then, problem **P3** is solved, or the close-form solution of the approximation problem **P4** is used to obtain weight $\boldsymbol{\gamma}$ (the reason for requiring only two backward passes for all $M+1$ modalities is explained in Appendix B). Next, the overall loss is calculated, and the model parameter is updated using an optimizer with learning rate η.

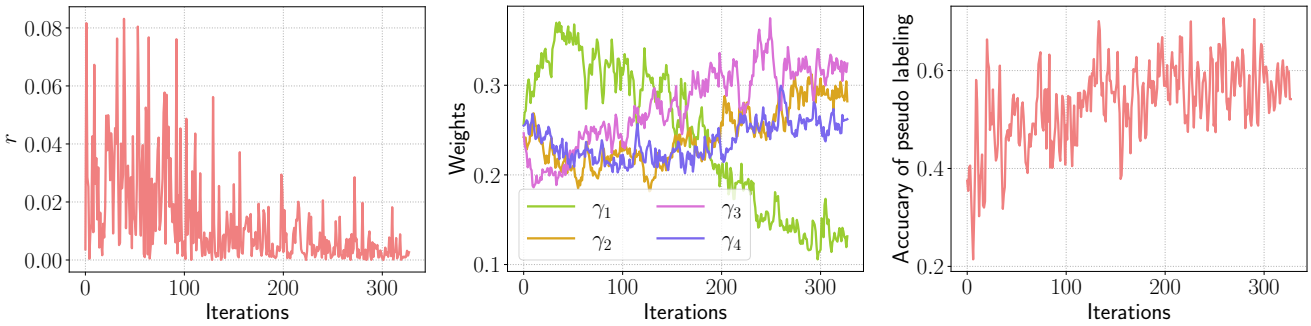
4 Numerical Results

Benchmark datasets: We evaluate our method on the multimodal emotion recognition task with two widely used benchmark datasets, IEMOCAP (Busso et al. 2008) and MSP-IMPROV (Busso et al. 2016). Both datasets contain acoustic, visual, and lexical modalities, corresponding to modalities 1, 2 and 3, respectively, in our model. IEMOCAP is composed of scripted and spontaneous dyadic conversations between actors. MSP-IMPROV features a wide range of spontaneous interactions, where participants engage in unscripted dialogues, providing rich and diverse data that reflect real-world communication scenarios. Following work (Zhao, Li, and Jin 2021), we select samples from the four classes — neutral, happy, sad and angry, to construct the datasets to be used in our experiments.

For each dataset, we split it evenly and randomly into two subsets. One subset is used directly as the source domain dataset, and the other, after some manipulations, serves as the target domain dataset. Specifically, for the samples in the target domain, we inject white noise into the acoustic modality with a signal-to-noise ratio (SNR) of 1.0; for the visual modality, the brightness of the video is reduced to 20 percent of its original level, and Gaussian noise with SNR$=$ 0.5 is added; for the lexical modality, 40 percent of the words in each utterance are randomly masked.

Baseline methods: In the following Comparison Studies section, we compare our model, Boomda, with DANN (Ganin et al. 2016), CDAN (Long et al. 2018), MADA (Pei et al. 2018), DALN (Chen et al. 2022a), PCL (Li et al. 2024b) and DADA(Ren et al. 2024), which are introduced in the Related Works section.

Implementation details: For the acoustic modality, WavLM (Chen et al. 2022b) followed by a TextCNN is employed as the feature encoder. For the visual modality, APViT pretrained on the RAF-DB (Li, Deng, and Du 2017) database is utilized for sequence feature extraction, and then a one-layer LSTM is utilized to encode the sequence feature. Bert-base (Devlin et al. 2018) and TextCNN are adopted for the lexical modality. The parameters in the last three layers of the pretrained models are set to be trainable, with all



(a) The characteristic (value r) of matrix $\boldsymbol{Q}$. (b) The balancing weight $\boldsymbol{\gamma}$. (c) The accuracy of pseudo labeling.

Figure 3: The training dynamics on the IEMOCAP dataset.

CA	PL	AL	AV	VL	AVL	ave.
✗	✗	30.99	39.34	36.37	40.31	36.75
✓	✗	46.19	44.74	43.42	47.85	45.55
✗	✓	42.53	44.58	**44.48**	48.78	45.09
✓	✓	45.96	45.88	39.91	53.36	46.27
✓	✓	**49.81**	**47.46**	42.83	**54.82**	**48.73**

Table 2: Results of ablation studies on dataset IEMOCAP. Symbol ✓ in the fourth row means correlation alignment is employed without balancing the modalities; that is, coefficients $\gamma_1 = \gamma_2 \cdots = \gamma_{M+1} = \frac{1}{M+1}$ are used for training.

other parameters frozen. The dimension of the representations $\boldsymbol{Z}_m, \forall m \in [M]$, is 256. The Adam optimizer is used for model training with learning rate 1×10^{-3}, momentum coefficient $(0.9, 0.999)$ and batch size 48. The hyperparameter settings are $\beta = 5 \times 10^{-4}$, $\alpha_1 = 0.5$, $\alpha_2 = 0.1$, and $M_v = 3$. More implementation details can be found from the appendix and the code in the supplementary material. Model performance is measured by the weighted F1 score, averaged over three runs on four 40GB Nvidia A40 GPUs.

Comparison Studies

The F1 score comparisons are reported in Table 1, where D.T. refers to direct transfer, meaning the model is trained with only the source domain data and directly tested on the target domain samples. In the table, "A", "V", and "L" represent acoustic, visual, and lexical modalities, respectively. We perform experiments with different combinations of modalities, and the "ave." columns correspond to the average results of these combinations. On average, Boomda achieves improvements over all compared models by at least 1.78 and 1.43, on the IEMOCAP and MSP-IMPROV datasets, respectively. More specifically, in all experiments, Boomda is only surpassed by other models in the VL setting on IEMOCAP and in the AV setting on MSP-IMPROV; except for these two settings, Boomda outperforms all other models. It is noteworthy that, without carefully balancing the modalities for domain alignment, some models can suffer from performance degradation with more modalities. For example, DALN achieves an F1 score of 47.23 with the AV (acoustic and visual) modalities on IEMOCAP, yet the F1 score decreases to 45.58 with the AVL modalities. A similar observation holds for DANN on the MSP-IMPROV dataset. With modality-balanced alignment, Boomda addresses this issue and demonstrates superior performance.

Ablation Studies

In the ablation studies, we analyze the effectiveness of the two primary designs: balanced correlation alignment (CA) and pseudo labeling (PL). As presented in Table 2, it is clear that balanced correlation alignment and pseudo labeling each can separately improve the F1 score by over 8.00 on average (from 36.77 to 45.55 and 45.09, respectively). When these two techniques are jointly adopted, the average F1 score further increases to 48.73. Comparing the last two rows, we conclude that the balanced alignment method in Boomda promotes the F1 score by approximately 2.50.

Moreover, we exhibit the value $r = \frac{\max\{|Q_{ij}| \mid i,j \in [M+1], i \neq j\}}{\min\{Q_{ii} \mid i \in [M+1]\}}$ (the ratio of the maximum absolute value of the off-diagonal entries to the minimum diagonal entry of matrix $\boldsymbol{Q}$) in Figure 3(a). It is demonstrated that r remains small throughout the training, which justifies that we can approximate matrix $\boldsymbol{Q}$ with its diagonal matrix $\tilde{\boldsymbol{Q}}$. Figure 3(b) illustrates how the balancing weight $\boldsymbol{\gamma}$ evolves during training: a relatively large weight is assigned to modality 1 (acoustic) at the early stage and then is reduced at the later stage. Figure 3(c) shows the accuracy of pseudo labeling in each training batch, which presents an increasing trend with the training process.

Due to the limited space, more experimental results, including ablation studies on MSP-IMPROV and running time comparisons, are provided in the supplementary material.

5 Conclusions

This paper investigates multimodal domain adaptation with a focus on balancing the alignment of different modalities. Based on the information bottleneck theory, we devise a model framework, in which each modality learns its representations and performs label prediction independently. Then, we propose a pseudo labeling strategy for target domain samples that exploits the predictions of all modalities. The target and source domains are aligned in the representation space using correlation alignment. To balance the alignment losses across modalities, we develop an efficient balanced multimodal domain adaptation approach, building upon a multi-objective optimization algorithm.

6 Acknowledgments

The work is supported by the National Key R&D Program of China (Grant No. 2024YFB4505602) and the National Natural Science Foundation of China (Grant No. 62306289).

References

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, Z.; Jiang, B.; and Jin, X. 2019. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3296–3303.

Chen, L.; Chen, H.; Wei, Z.; Jin, X.; Tan, X.; Jin, Y.; and Chen, E. 2022a. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7181–7190.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Désidéri, J.-A. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6): 313–318.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, C.; Li, T.; Liu, Y.; Wen, Z.; Hua, T.; Wang, Y.; and Zhao, H. 2021. Improving Multi-Modal Learning with Uni-Modal Teachers. *arXiv e-prints*, arXiv–2106.

Dua, D.; Strubell, E.; Singh, S.; and Verga, P. 2023. To Adapt or to Annotate: Challenges and Interventions for Domain Adaptation in Open-Domain Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14429–14446.

Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20029–20038.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Hu, H.; Luan, Y.; Chen, Y.; Khandelwal, U.; Joshi, M.; Lee, K.; Toutanova, K.; and Chang, M.-W. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12065–12075.

Hu, Y.; Xian, R.; Wu, Q.; Fan, Q.; Yin, L.; and Zhao, H. 2024. Revisiting scalarization in multi-task learning: A theoretical perspective. *Advances in Neural Information Processing Systems*, 36.

Kawaguchi, K.; Deng, Z.; Ji, X.; and Huang, J. 2023. How Does Information Bottleneck Help Deep Learning? In *International Conference on Machine Learning*. PMLR.

Kennerley, M.; Wang, J.-G.; Veeravalli, B.; and Tan, R. T. 2023. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11484–11493.

Kruse, C. N.; Hansen, L.; and Heinrich, M. P. 2021. Multimodal unsupervised domain adaptation for deformable registration based on maximum classifier discrepancy. In *Bildverarbeitung für die Medizin 2021: Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021*, 192–197. Springer.

Lee, J.; Das, D.; Choo, J.; and Choi, S. 2023. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16380–16389.

Li, J.; Yu, Z.; Du, Z.; Zhu, L.; and Shen, H. T. 2024a. A Comprehensive Survey on Source-free Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01): 1–22.

Li, J.; Zhang, Y.; Wang, Z.; Hou, S.; Tu, K.; and Zhang, M. 2024b. Probabilistic contrastive learning for domain adaptation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 1001–1009.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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, W.; Luo, Z.; Cai, Y.; Yu, Y.; Ke, Y.; Junior, J. M.; Gonçalves, W. N.; and Li, J. 2021. Adversarial unsupervised domain adaptation for 3D semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176: 211–221.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Munro, J.; and Damen, D. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 122–132.

Navon, A.; Shamsian, A.; Achituve, I.; Maron, H.; Kawaguchi, K.; Chechik, G.; and Fetaya, E. 2022. Multi-Task Learning as a Bargaining Game. In *International Conference on Machine Learning*, 16428–16446. PMLR.

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

Ren, L.; Chen, C.; Wang, L.; and Hua, K. 2024. Towards improved proxy-based deep metric learning via data-augmented domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14811–14819.

Saxe, A. M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B. D.; and Cox, D. D. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124020.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shi, J.; Shang, C.; Sun, Z.; Yu, L.; Yang, X.; and Yan, Z. 2024. PASSION: Towards Effective Incomplete Multi-Modal Medical Image Segmentation with Imbalanced Missing Rates. *arXiv preprint arXiv:2407.14796*.

Singhal, P.; Walambe, R.; Ramanna, S.; and Kotecha, K. 2023. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11: 6973–7020.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 443–450. Springer.

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, J.; Zhang, X.; Han, S.; Ruan, Y.-P.; and Li, T. 2024. RedCore: Relative Advantage Aware Cross-Modal Representation Learning for Missing Modalities with Imbalanced Missing Rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15173–15182.

Sun, J.; Zhang, X.; Hong, S.; Zhu, J.; and Zeng, L. 2025. Adversarial Alignment with Anchor Dragging Drift (A3D2): Multimodal Domain Adaptation with Partially Shifted Modalities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 19680–19690.

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 Jia, K. 2020. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5940–5947.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Wang, M.; and Deng, W. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153.

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, W.; Zhong, Z.; Wang, W.; Chen, X.; Ling, C.; Wang, B.; and Sebe, N. 2023. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24090–24099.

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.

Xing, B.; Ying, X.; Wang, R.; Yang, J.; and Chen, T. 2023. Cross-modal contrastive learning for domain adaptation in 3D semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2974–2982.

Yadav, N.; Alam, M.; Farahat, A.; Ghosh, D.; Gupta, C.; and Ganguly, A. R. 2023. CDA: Contrastive-adversarial Domain Adaptation. *arXiv preprint arXiv:2301.03826*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

Zhang, X.; Sun, J.; Hong, S.; and Li, T. 2024. Amanda: Adaptively modality-balanced domain adaptation for multi-modal emotion recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, 14448–14458.

Zhao, H.; Zhang, S.; Wu, G.; Moura, J. M.; Costeira, J. P.; and Gordon, G. J. 2018. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31.

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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2608–2618.

Supplementary Material of Paper "Boomda: Balanced Multi-objective Optimization for Multimodal Domain Adaptation".

Appendix

The numbers of figures and tables, and the references in the appendix follow those in the paper.

A. Proof of Theorem 1

Proof. We first rewrite problem **P4** as follows.

$$\mathbf{P4}: \min_{\boldsymbol{\gamma}} \ \boldsymbol{\gamma}^T \tilde{\boldsymbol{Q}} \boldsymbol{\gamma} \tag{15a}$$

$$\text{s.t.} \ \ \boldsymbol{\gamma} \geq \mathbf{0} \tag{15b}$$

$$\mathbf{1}^T \cdot \boldsymbol{\gamma} = 1. \tag{15c}$$

Let $\boldsymbol{\mu}$ and λ denote the Lagrangian multipliers associated with constraints (15b) and (15c), respectively. Then the Lagrangian function of problem **P4** writes:

$$\Gamma(\boldsymbol{\gamma}, \boldsymbol{\mu}, \lambda) = \boldsymbol{\gamma}^T \tilde{\boldsymbol{Q}} \boldsymbol{\gamma} - \boldsymbol{\mu}^T \boldsymbol{\gamma} + \lambda(\mathbf{1}^T \cdot \boldsymbol{\gamma} - 1). \tag{16}$$

Then, we have the KKT conditions of problem **P4**:

$$\frac{\partial \Gamma(\boldsymbol{\gamma}, \boldsymbol{\mu}, \lambda)}{\partial \boldsymbol{\gamma}} = \tilde{\boldsymbol{Q}} \boldsymbol{\gamma} - \boldsymbol{\mu} + \lambda \mathbf{1} = 0; \tag{17a}$$

$$\frac{\partial \Gamma(\boldsymbol{\gamma}, \boldsymbol{\mu}, \lambda)}{\partial \lambda} = \mathbf{1}^T \cdot \boldsymbol{\gamma} - 1 = 0; \tag{17b}$$

$$\boldsymbol{\gamma} \geq 0; \tag{17c}$$

$$\boldsymbol{\mu} \geq 0; \tag{17d}$$

$$\boldsymbol{\mu}^T \boldsymbol{\gamma} = 0. \tag{17e}$$

We first show that $\boldsymbol{\mu} = \mathbf{0}$ holds. We will prove it by contradiction in the following.

Suppose that there exists an element μ_m in $\boldsymbol{\mu}$ such $\mu_m > 0$. Then according to (17e), the corresponding element γ_m satisfies:

$$\gamma_m = 0. \tag{18}$$

Considering that $\tilde{\boldsymbol{Q}}$ is a diagonal matrix, we can derive from (17a) that $\lambda = \mu_m > 0$. Then for any $m' \in [M+1]$ and $m' \neq m$, we have

$$\mu_{m'} = \tilde{Q}_{m'} \gamma_{m'} + \lambda > 0, \tag{19}$$

where $\tilde{Q}_{m'}$ denote the m'-th diagonal entry of $\tilde{\boldsymbol{Q}}$.

From Eqs. (17e) and (19), one can obtain that

$$\gamma_{m'} = 0, \forall m' \in [M+1] \text{ and } m' \neq m. \tag{20}$$

Combining Eqs. (18) and (20) leads to $\boldsymbol{\gamma} = \mathbf{0}$, which contradicts with (17b) and thus concludes that $\boldsymbol{\mu} = \mathbf{0}$ holds.

Upon plugging $\boldsymbol{\mu} = \mathbf{0}$ into Eq. (17a), we attain:

$$\boldsymbol{\gamma} = \lambda \tilde{\boldsymbol{Q}}^{-1} \mathbf{1}. \tag{21}$$

Combining Eqs. (21) and (17b) yields the desired result:

$$\boldsymbol{\gamma} = \frac{\tilde{\boldsymbol{Q}}^{-1} \mathbf{1}}{\mathbf{1}^T \tilde{\boldsymbol{Q}}^{-1} \mathbf{1}},$$

which finishes the proof. □

B. Derivation of Problem P2 from Problem P1

As per works (Désidéri 2012; Sener and Koltun 2018), we adopt multiple gradient descent algorithm (MGDA) to solve problem (12). MGDA is built upon the Karush-Kuhn-Tucker (KKT) conditions, which are necessary conditions for the solution to be Pareto optimal and defines the Pareto stationary point. Specifically, any solution $\boldsymbol{\theta}$ is called a Pareto stationary point if there exists a vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_{M+1}]^T$, such that the following conditions are satisfied: a) $\boldsymbol{\gamma} \geq \mathbf{0}$, b) $\mathbf{1}^T \cdot \boldsymbol{\gamma} = 1$, and c) $\sum_{m \in [M+1]} \gamma_m \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{1}$ and $\mathbf{0}$ represent all-one and all-zero vector of proper size, respectively.

With these KKT conditions, an optimization problem is formulated as follows.

$$\mathbf{P1}: \min_{\boldsymbol{\gamma}} \ || \sum_{m \in [M+1]} \gamma_m \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) ||_2^2$$
$$s.t. \ \ \boldsymbol{\gamma} \geq \mathbf{0}, \ \ \mathbf{1}^T \cdot \boldsymbol{\gamma} = 1.$$

Suppose that $\boldsymbol{\gamma}^* = [\gamma_1^*, \gamma_2^*, \cdots, \gamma_{M+1}^*]$ is the optimal solution of problem **P1**, then two outcomes follows according to the resultant $\sum_{m \in [M+1]} \gamma_m^* \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta})$. 1) $\sum_{m \in [M+1]} \gamma_m^* \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) = \mathbf{0}$, which means that the corresponding $\boldsymbol{\theta}$ is a Pareto stationary point, and 2) $\sum_{m \in [M+1]} \gamma_m^* \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) \neq \mathbf{0}$, which is a descent direction for all objectives, and can be applied to update the model parameters; equivalently, objective $\sum_{m \in [M+1]} \gamma_m^* \mathcal{L}_m^{CA}(\boldsymbol{\theta})$ is optimized.

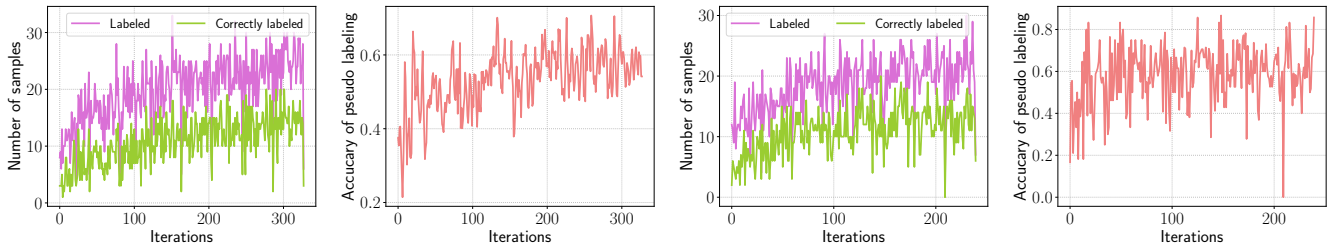
Following work of (Sener and Koltun 2018), we can have an upper bound of the objective in problem **P1**:

$$\begin{aligned} & || \sum_{m \in [M+1]} \gamma_m \nabla_{\boldsymbol{\theta}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) ||_2^2 \\ = & || \sum_{m \in [M+1]} \gamma_m \frac{\partial \boldsymbol{Z}_{M+1}}{\partial \boldsymbol{\theta}^e} \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) ||_2^2 \\ \leq & || \frac{\partial \boldsymbol{Z}_{M+1}}{\partial \boldsymbol{\theta}^e} ||_F^2 \cdot || \sum_{m \in [M+1]} \gamma_m \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) ||_2^2. \end{aligned} \tag{22}$$

We can replace the objective of problem **P1** using its upper bound with term $|| \frac{\partial \boldsymbol{Z}_{M+1}}{\partial \boldsymbol{\theta}^e} ||_F^2$ dropped, since $|| \frac{\partial \boldsymbol{Z}_{M+1}}{\partial \boldsymbol{\theta}^e} ||_F^2$ is independent from $\boldsymbol{\gamma}$. $\nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}), \forall m \in [M+1]$ can be calculated with just two backward passes, one for $\nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_{M+1}^{CA}(\boldsymbol{\theta})$, and the other for $\nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}), \forall m \in [M]$. This is because that $\nabla_{\boldsymbol{Z}_{m'}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) = \mathbf{0}, \forall m, m' \in [M]$, and $m \neq m'$, implying that the gradient $\nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}), \forall m \in [M]$ can be obtained via computing the gradient of $\sum_{m \in [M]} \mathcal{L}_m^{CA}(\boldsymbol{\theta})$ with respect to $\boldsymbol{Z}_{M+1}$.

With the above analysis, we can have an efficient alternative of problem **P1**:

$$\mathbf{P2}: \min_{\boldsymbol{\gamma}} \ || \sum_{m \in [M+1]} \gamma_m \nabla_{\boldsymbol{Z}_{M+1}} \mathcal{L}_m^{CA}(\boldsymbol{\theta}) ||_2^2$$
$$\text{s.t.} \ \ \boldsymbol{\gamma} \geq \mathbf{0}, \ \ \mathbf{1}^T \cdot \boldsymbol{\gamma} = 1.$$



(a) Number of labeled samples. (b) Accuracy of labeling. (c) Number of labeled samples. (d) Accuracy of labeling.

Figure 4: The pseudo labeling during training ((a) and (b) are the results on the IEMOCAP dataset; (c) and (d) are the results on the MSP-IMPROV dataset).

CA	PL	AL	AV	VL	AVL	ave.
✗	✗	25.09	34.48	35.77	38.76	33.53
✓	✗	26.73	40.80	41.69	42.43	37.91
✗	✓	31.12	39.91	42.61	47.21	40.21
✓̂	✓	30.79	**42.69**	42.97	44.68	40.28
✓	✓	**33.30**	40.31	**45.10**	**47.29**	**41.50**

Table 3: Results of ablation studies on the MSP-IMPROV dataset. Symbol ✓̂ in the fourth row means correlation alignment is employed without balancing the modalities.

C. Implementation Details

Preprocess of threee modalities: The audio is sampled at frequency 16kHz with a receptive field of 25ms and a frame shift of 20ms. The video is first processed with MTCNN (Zhang et al. 2016) to obtain aligned faces and each frame is cropped to size of 112 × 112. For each video utterance, we evenly sample 8 frames as the raw feature of the visual modality. For the text of the target domain, all the masked words are replaced with word "MASK".

D. Ablation Studies

Ablation Studies on the MSP-IMPROV Dataset: The ablation studies on the MSP-IMPROV dataset yield results similar to those observed on IEMOCAP. Namely, balanced correlation alignment and pseudo labeling independently enhance F1 score by an average of over 4.00 and 7.00, respectively, increasing it from 33.53 to 37.91 and 40.21. When both techniques are applied simultaneously, the average F1 score is further improved to 41.50. Comparing the last two rows, it is evident that the balanced alignment method in Boomda results in an additional improvement of approximately 1.20.

Pseudo Labeling: Figure 4(a) illustrates the number of labeled samples and the number of correctly labeled samples during training on the IEMOCAP dataset. As training progresses, more samples are selected for labeling, indicating increased agreement between modalities. Concurrently, the accuracy of labeling improves over iterations, as shown in Figure 4(b). Similar trends are observed with the MSP-IMPROV dataset, as depicted in Figures 4(c) and (d).

	IEMOCAP	MSP-IMPROV
MGDA	137s	118s
Boomda	116s	94s

Table 4: Comparisons of running time per epoch.

E. Running Time Comparisons

The original MGDA algorithm leverages Frank-Wolfe algorithm to solve problem **P3** at each iteration. In comparison, Boomda adopts the closed-form solution of problem **P4** as the balancing weight γ. Therefore, Boomda is computationally more efficient than MGDA, which is validated by the running time per epoch reported in Table 4. Specifically, Boomda reduces therunning time of each epoch by about 15% and 20% for datasetsIEMOCAP and MSP-IMPROV, respectively.