SEAL: Separate and Augment with Pseudo-Labeling for Efficient Multimodal Multi-Target Domain Adaptation

Anonymous ACL submission

Abstract

This paper investigates the multimodal multi-target domain adaptation problem where independent shifts of all modalities lead to an exponential number of multimodal target domains. We categorize them into F-target domains, where only one modality shifts, and U-target domains, where multiple modalities shift simultaneously. To alleviate the burden of collecting data from all domains, we propose a novel multimodal multi-target domain adaptation approach that requires only labeled samples from the source domain and unlabeled samples from F-target domains, thus achieving linear sample complexity. Specifically, we first disentangle each modality's representation into task-relevant and domain-relevant components via mutual information maximization. Then, we augment source domain samples by recombining these components to emulate labeled samples from F-target and U-target domains. Moreover, we introduce a pseudo-labeling strategy that exploits the unshifted modalities of each F-target domain sample to generate pseudo labels for training. The overall design follows the principle of "**SE**parate and **A**ugment with pseudo-**L**abeling" (**SEAL**) to enable efficient multimodal multi-target domain adaptation. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches on widely used benchmark datasets. The code is available in the supplementary material.

1 Introduction

Multimodal learning has demonstrated superior performance over unimodal approaches by exploiting complementary information across visual, acoustic, and lexical modalities (Yuan et al., 2025), leading to its widespread adoption in applications ranging from autonomous driving (Chen et al., 2025) to biomedicine (Huang et al., 2025). However, multimodal models typically rely on data collected from heterogeneous devices and distinct physical environments, where different modalities reside in disparate spatial and sensory contexts. As a result, multimodal learning faces two key challenges: the high cost of collecting and annotating multimodal data, and increased vulnerability to domain shifts, since distributional changes in any modality can severely degrade overall performance. Domain adaptation addresses these challenges by transferring knowledge from a labeled source domain to an unlabeled target domain (Zhu et al., 2023).

A \ V	0	1	2
0	S(0,0)	F(0,1)	F(0,2)
1	F(1,0)	U(1,1)	U(1,2)
2	F(2,0)	U(2,1)	U(2,2)

(a) All domains

Domain type	Number of domains
Source domain	1
F-target domains	$M(D-1)$
U-target domains	$D^M - M(D-1) - 1$
All domains	D^M

(b) Number of domains

Figure 1: All domains & Number of domains

In this paper, we investigate domain adaptation in a practical yet underexplored setting: multimodal learning with multiple target domains. In real-world scenarios, each modality may experience multiple domain shifts (e.g., visual data captured under sunny, rainy, or foggy conditions). Formally, unsupervised multi-target domain adaptation assumes labeled data from a single source domain and unlabeled data from multiple target domains (Isobe et al., 2021).

This problem becomes substantially more challenging in multimodal settings, where the number of domains grows exponentially with the number of modalities. Specifically, with M modalities and D unimodal domains per modality, the total number of multimodal domains is D^M, making comprehensive data collection impractical. For illustration, we consider $M = 2$ modalities (e.g., visual and acoustic) and $D = 3$ domains per modality, resulting in nine multimodal domains (Figure 1(a)). These domains are categorized into: a labeled source domain, $M(D-1)$ F-target domains with unlabeled observations, and $(D^M - M(D-1) - 1)$ U-target do-

mains that are entirely unobserved, as summarized in Figure 1(b).

The objective of multimodal multi-target domain adaptation in this work is to leverage labeled data from the source domain and unlabeled data from the F-target domains to learn a model that generalizes to both F-target and unobserved U-target domains. In contrast to conventional domain adaptation methods that require unlabeled data from all target domains, our approach significantly reduces the sample burden: the number of required domains scales linearly with the number of modalities, rather than exponentially. This efficiency arises from the structure of the F-target domains, where each modality independently undergoes all unimodal domain shifts while the remaining modalities are fixed in their source domains. Consequently, although U-target domains are never observed during training, the model is exposed to all unimodal shift patterns, enabling robust generalization to unseen multimodal domains.

Current mainstream domain adaptation methods mitigate domain shifts by learning domain-invariant representations that generalize across source and target domains (Stojanov et al., 2021), typically discarding domain-specific information while preserving task-relevant features. However, the majority of existing approaches are designed for single-source, single-target, and unimodal settings (Li et al., 2024a), with limited extensions to multi-source or multi-target scenarios (Pei et al., 2024; Isobe et al., 2021). In the multimodal multi-target setting considered in this work, directly aligning the source domain with all $(D^M - 1)$ target domains becomes significantly more challenging and can lead to suboptimal performance.

Different from explicit domain alignment, we propose to enhance robustness by exposing the task classifier to diverse domain variations during training. Specifically, we learn modality-specific representations and decompose them into task-relevant and domain-relevant components via mutual information maximization (Poole et al., 2019; Wang et al., 2023). We then augment source-domain samples in the representation space by combining their task-relevant components with domain-relevant components drawn from F-target and U-target domains. This process synthesizes target-domain-like representations while preserving source labels, enabling supervised training without requiring labeled target-domain data. By emulating labeled target samples in the representation space, we enrich the domain diversity of the training data.

In addition, we develop a pseudo-labeling strategy that exploits the characteristic of F-target domain samples, where only one modality undergoes shift while the others remain in their source domains. For each F-target sample, task labels are independently predicted using the stable modalities and then aggregated to form a pseudo label. To enhance the reliability of supervision, samples with low-entropy (i.e., high-confidence) pseudo labels are assigned larger weights during training.

In summary, this paper proposes a novel approach that features **SE**parating and **A**ugmenting with pseudo-**L**abeling, dubbed **SEAL**, for efficient multimodal multi-target domain adaptation. The primary contributions are threefold:

- We propose a mutual-information-based representation disentanglement and augmentation method for multimodal multi-target domain adaptation. This method mimics unobserved target domains samples, thereby reducing the sample complexity from exponential to linear.
- We generate pseudo labels for F-target samples using modalities that remain in their source domains. These pseudo-labeled samples are incorporated into training with confidence-based weights determined by prediction entropy.
- Extensive empirical results demonstrate the effectiveness of the proposed method, and show that SEAL outperforms SOTA approaches.

2 Related Works

2.1 Domain adaptation approaches

The majority of current domian approaches is developed for single source, single target, and single modality settings, as comprehensively reviewed in survey papers (Wang and Deng, 2018; Li et al., 2024a) . In the following, we selectively introduce two branches: moment based and adversarial based approaches. Moment based methods mitigate domain shifts by minimizing moment-based distribution discrepancy across domains. Maximum mean discrepancy (MMD)-based approaches, such as DDC (Tzeng et al., 2014) and MK-MMD (Long et al., 2015), align first-order statistics by matching feature means. CORAL (Sun et al., 2016; Sun and Saenko, 2016) and JDDA (Chen et al., 2019) extend this idea to second-order statistics by aligning feature covariances, while CMD (Zellinger

et al., 2017) further generalizes moment matching to higher-order central moments.

Adversarial learning was first introduced to domain adaptation by DANN (Ganin et al., 2016), which inspired a broad class of adversarial-based methods. CDAN (Long et al., 2018) and CAN (Wu et al., 2021) additionally introduce label information to enable conditional adversarial alignment. Contrastive learning has also been integrated to achieve class-level alignment, as in CDA (Yadav et al., 2023) and LUHP (Zhang et al., 2024). Beyond standard adversarial frameworks, DALN (Chen et al., 2022a) proposes a discriminator-free approach by reusing the task classifier, while f-DD (Wang and Mao, 2024) introduces an f-domain discrepancy with theoretical guarantees on target error and sample complexity. Recent methods further enhance adversarial adaptation through data augmentation, such as PCL (Li et al., 2024b) in the raw feature space and DADA (Ren et al., 2024) in the representation space. In multimodal settings, A^3D^2 (Sun et al., 2025) incorporates optimal transport for cross-modality alignment.

2.2 Multi-source/multi-target domain adaptation

Building on earlier domain adaptation techniques, recent work on multi-source and multi-target domain adaptation addresses domain shifts through alignment and knowledge transfer. In multi-source settings, representative approaches explore adversarial alignment across domains (Lin et al., 2020; Jiang et al., 2025) or construct pseudo target domains to bridge distribution gaps (Ren et al., 2022). Source-free adaptation further extends this line by leveraging uncertainty modeling and contrastive learning to transfer knowledge from pretrained models without access to source data (Pei et al., 2024; Zhao et al., 2025).

For multi-target domain adaptation, existing methods mainly focus on disentangling domain-shared and domain-specific representations (Gholami et al., 2020), curriculum-based adaptation across targets (Roy et al., 2021), and expert alignment with knowledge distillation to improve generalization (Isobe et al., 2021; Saporta et al., 2021). Graph-based modeling has been introduced to capture structured cross-domain semantic relationships (Yang et al., 2022).

Existing works on multimodal domain adaptation primarily focus on single-source and single-target settings (Dong et al., 2025), with only limited exploration of multi-source scenarios (Zhao et al., 2025). In contrast, the multimodal multi-target setting considered in this paper has received little to no research attention.

3 Method: SEAL

Before introducing our method, SEAL, we first define the notations to be used. To improve readability, we provide a notation table in the appendix.

Modalities and raw features: Suppose in a multimodal classification task, each sample comprises M modalities, represented by the set $\{\boldsymbol{x}_m\}_{m=1}^M$, where $\boldsymbol{x}_m \in \mathbb{R}^{p_m}$ denotes the raw feature of modality m. For simplicity, we define an auxiliary $(M+1)$-th modality that concatenates all modality features: $\boldsymbol{x}_{M+1} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M]$.

Domains: Each modality $m \in \{1, 2, \cdots, M\}$ is associated with D_m unimodal domains, indexed by $d_m \in \{0, 1, \cdots, D_m - 1\}$, where $d_m = 0$ indicates the source domain and $d_m \geq 1$ refers to target domains. We define a multimodal domain indicator for each sample as $\boldsymbol{d} := [d_1, d_2, \cdots, d_M]$, where $d_m \in \{0, 1, \cdots, D_m - 1\}$ specifies which domain modality m is from. With the above notations, we can use $\mathcal{D}_S = \{\boldsymbol{d} | d_1 = d_2 = \cdots = d_M = 0\}$, $\mathcal{D}_F = \{\boldsymbol{d} | \sum_{m=1}^M \mathbf{1}_{(d_m \neq 0)} = 1\}$ and $\mathcal{D}_U = \{\boldsymbol{d} | \sum_{m=1}^M \mathbf{1}_{(d_m \neq 0)} \geq 2\}$ to represent the set of source domain, F-target domains and U-target domains, respectively. Assume that there are N_S labeled samples from the source domain, and $N_F/|\mathcal{D}_F|$ unlabeled samples from each F-target domain, where $|\cdot|$ means the cardinality of a set (this uniformity assumption is for notational simplicity, and real-world datasets may vary in sample counts per domain).

Samples: Let $\boldsymbol{x}_n^{\boldsymbol{d}} := \{\boldsymbol{x}_{m,n}^{d_m}\}_{m=1}^M, n \in \{1, 2, \cdots, N_S + N_F\}$ denote the feature of n-th sample from domain $\boldsymbol{d}$, where $\boldsymbol{x}_{m,n}^{d_m}$ represents the feature of modality m from domain d_m. If the sample is from the source domain, it is annotated with label $\boldsymbol{y}_n := \{\boldsymbol{y}_{m,n}\}_{m=1}^{M+1}$, where $\boldsymbol{y}_{m,n}$ 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. The training set consists of: 1) labeled source domain samples, $\{\{\boldsymbol{x}_n^{\boldsymbol{d}}, \boldsymbol{y}_n\}_{n=1}^{N_S} | \boldsymbol{d} \in \mathcal{D}_S\}$; and 2) unlabeled F-target domain samples, $\{\{\boldsymbol{x}_n^{\boldsymbol{d}}\}_{n=1}^{N_F} | \boldsymbol{d} \in \mathcal{D}_F\}$.

Random variables: Let $\boldsymbol{X}_m$ and $\boldsymbol{Y}$ represent general feature and label random variables for modality m, with $\boldsymbol{x}_{m,n}$ and y_n being their specific in-

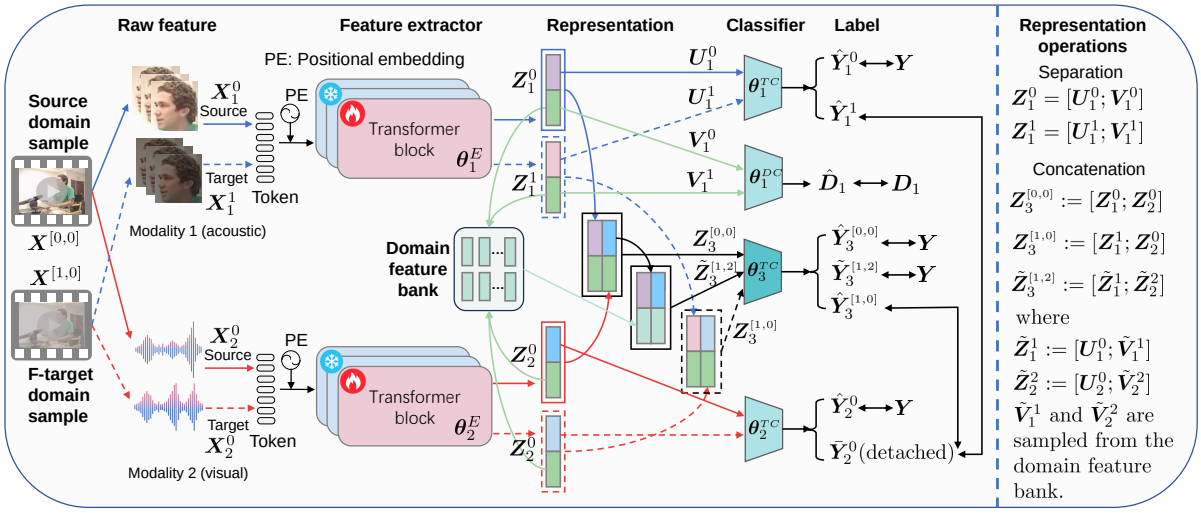

Figure 2: Model framework with 2 modalities as an example (multimodal representation $\boldsymbol{Z}_3$ is a concatenation of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$; solid and dashed regular arrows represent the flows of source and F-target domains, respectively; double-headed arrows represent supervision signals. Blue, red and green arrows denote data flow of modality 1, modality 2 and domain feature, respectively; for succinctness, we omit the domain classifier for modality 2.

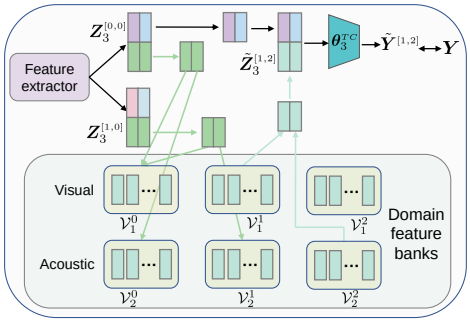

Figure 3: The domain feature bank (upper panel) and information flow of representation learning (lower panel)

stances. Let $\boldsymbol{Z}_m \in \mathbb{R}^p$, derived from $\boldsymbol{X}_m$, denote the learned representation of modality m, with instance $\boldsymbol{z}_{n,m}$. For consistency, we use $\boldsymbol{X}_{M+1}$ and $\boldsymbol{Z}_{M+1}$ to denote the aggregated multimodal feature and its corresponding representation, respectively. When necessary, superscripts d_m and $\boldsymbol{d}$ are used for the $\boldsymbol{X}_m$ and $\boldsymbol{Z}_m$, $m \in [M+1]$ to indicate the domain origin of features and representations.

3.1 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 are tunable and the bottom layers are frozen during training. Following the pretrained models are sequence encoders (omitted in Figure 2 and will be specified in the Numerical Results section) 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 or stochastic encoder function $p(\boldsymbol{Z}_m|\boldsymbol{X}_m; \boldsymbol{\theta}_m^{EN}) : \mathbb{R}^{p_m} \to \mathbb{R}^p$ with trainable parameter $\boldsymbol{\theta}_m^{EN}$; then we have $\boldsymbol{Z}_m \sim p(\boldsymbol{Z}_m|\boldsymbol{X}_m; \boldsymbol{\theta}_m^{EN})$.

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. The multimodal representation is then fed to the task classifier $q(\boldsymbol{Y}|\boldsymbol{Z}_{M+1}; \boldsymbol{\theta}_{M+1}^{TC})$ with parameter $\boldsymbol{\theta}_{M+1}^{TC}$ for label prediction; that is, the multimodal model outputs the probability distribution over all classes as: $\hat{\boldsymbol{Y}}_{M+1} \sim q(\boldsymbol{Y}|\boldsymbol{Z}_{M+1}; \boldsymbol{\theta}_{M+1}^{TC})$. The above multimodal model framework is representative in the field of multimodal learning, and our multimodal domain adaptation is developed upon this framework.

As Figure 2 illustrates, in addition to the multimodal task classifier, we introduce a unimodal task classifier and a domain classifier for each individual modality (the domain classifier for modality 2 is omitted in the figure). The details of these classifiers will be elaborated in the sequel.

B. Representation Separation. It is natural to assume that each unimodal representation $\boldsymbol{Z}_m$ can be decomposed into a task-relevant component $\boldsymbol{U}_m$

and a domain-relevant component $\boldsymbol{V}_m$, such that $\boldsymbol{Z}_m = [\boldsymbol{U}_m; \boldsymbol{V}_m]$. The former captures information for the classification task, while the latter holds domain specific information. However, without further constraints, task and domain information remain entangled in $\boldsymbol{Z}_m$, and cannot be directly separated into $\boldsymbol{U}_m$ and $\boldsymbol{V}_m$. To address this, we propose to explicitly disentangle $\boldsymbol{Z}_m$ using two complementary strategies: 1) mutual information maximization and 2) dimension-wise decorrelation.

Mutual information maximization: As shown in Figure 2, to extract task-relevant information, we introduce a task classifier $q(\boldsymbol{Y}|\boldsymbol{U}_m; \boldsymbol{\theta}_m^{TC})$, with parameter, $\boldsymbol{\theta}_m^{TC}$, to approximate the true classifier $p(\boldsymbol{Y}|\boldsymbol{U}_m)$. Encouraging $\boldsymbol{U}_m$ to retain task-specific information is equivalent maximizing the mutual information between $\boldsymbol{U}_m$ and the corresponding label $\boldsymbol{Y}$, $I(\boldsymbol{U}_m, \boldsymbol{Y})$, which can be written as:

$$\begin{aligned} I(\boldsymbol{U}_m, \boldsymbol{Y}) :=& \mathbb{E}_{p(\boldsymbol{U}_m, \boldsymbol{Y})} \left[\log \frac{p(\boldsymbol{Y}|\boldsymbol{U}_m)}{p(\boldsymbol{Y})} \right] \\ =& \mathbb{E}_{p(\boldsymbol{U}_m, \boldsymbol{Y})} \left[\log \frac{q(\boldsymbol{Y}|\boldsymbol{U}_m)}{p(\boldsymbol{Y})} \cdot \frac{p(\boldsymbol{Y}|\boldsymbol{U}_m)}{q(\boldsymbol{Y}|\boldsymbol{U}_m)} \right] \\ =& \mathbb{E}_{p(\boldsymbol{U}_m, \boldsymbol{Y})} \left[\log q(\boldsymbol{Y}|\boldsymbol{U}_m) \right] + H(\boldsymbol{Y}) \\ &+ \mathbb{E}_{p(\boldsymbol{U}_m)} \left[D_{KL} \left(p(\boldsymbol{Y}|\boldsymbol{U}_m) || q(\boldsymbol{Y}|\boldsymbol{U}_m) \right) \right] \\ \geq& \mathbb{E}_{p(\boldsymbol{U}_m, \boldsymbol{Y})} \left[\log q(\boldsymbol{Y}|\boldsymbol{U}_m) \right] + H(\boldsymbol{Y}), \end{aligned} \tag{1}$$

where $D_{KL}(\cdot||\cdot)$ denotes the KL divergence of two distributions, and it is non-negative; $H(\cdot)$ represents the entropy of a random variable. Note that the task label $\boldsymbol{Y}$ is available only for source domain samples, and thus the mutual information maximization for $\boldsymbol{U}_m$ and $\boldsymbol{Y}$ is performed over the source domain. For given dataset, the entropy of the label is a constant independent of the model parameters. Thus, we just need to focus on maximizing the first term in Eq. (1), which is equivalent to minimizing the following loss function for all $m \in \{1, 2, \cdots, M+1\}$:

$$\begin{aligned} \mathcal{L}_m^{\text{TAS}}(\boldsymbol{\theta}) :=& - \mathbb{E}_{p(\boldsymbol{U}_m, \boldsymbol{Y})} \left[\log q(\boldsymbol{Y}|\boldsymbol{U}_m) \right] \\ =& - \frac{1}{N_S} \sum_{n=1}^{N_S} \sum_{c=1}^{C} [\boldsymbol{y}_n]_c \log [\hat{\boldsymbol{y}}_{m,n}]_c. \end{aligned} \tag{2}$$

Eq.(2) corresponds to the standard categorical cross-entropy loss of classifier $\hat{\boldsymbol{Y}} \sim q(\boldsymbol{Y}|\boldsymbol{U}_m; \boldsymbol{\theta}_m^{TC})$.

Similarly, each modality is associated with a domain classifier $q(\boldsymbol{Y}|\boldsymbol{U}_m; \boldsymbol{\theta}_m^{TC})$, with parameter $\boldsymbol{\theta}_m^{TC}$. We have the lower bound for the mutual information of $\boldsymbol{V}_m$ and $\boldsymbol{D}_m$ as follows.

$$I(\boldsymbol{V}_m, \boldsymbol{D}_m) \geq \mathbb{E}_{p(\boldsymbol{V}_m, \boldsymbol{D}_m)} \left[\log q(\boldsymbol{D}_m|\boldsymbol{V}_m) \right] + H(\boldsymbol{D}_m), \tag{3}$$

where $H(\boldsymbol{D}_m)$ is a constant, and thus we focus on maximizing the first term. All samples in the source and F-target domains are associated with a domain label. Hence, maximizing the above lower bound is equivalent to minimizing the following loss function:

$$\begin{aligned} \mathcal{L}_m^{\text{DOM}}(\boldsymbol{\theta}) :=& - \mathbb{E}_{p(\boldsymbol{V}_m, \boldsymbol{D}_m)} \left[\log q(\boldsymbol{D}_m|\boldsymbol{V}_m) \right] \\ =& - \frac{1}{N_S + N_F} \sum_{n=1}^{N_S+N_F} \sum_{i=1}^{D_m} [d_{m,n}]_i \log [\hat{d}_{m,n}]_i, \end{aligned} \tag{4}$$

where we abuse the scalar domain index (or label) $d_{m,n}$ and $\hat{d}_{m,n}$ as their corresponding one-hot vector.

The information flow of representation is exhibited in the lower panel of Figure 3, and the loss for representation separation is as follows.

$$\mathcal{L}^{\text{SEP}}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \mathcal{L}_m^{TAS}(\boldsymbol{\theta}) + \mathcal{L}_m^{DOM}(\boldsymbol{\theta}). \tag{5}$$

Dimension-wise decorrelation: While mutual information maximization helps encode task-relevant information in $\boldsymbol{U}_m$ and domain-relevant information in $\boldsymbol{V}_m$, it does not guarantee complete disentanglement; that is, task information may still leak into $\boldsymbol{V}_m$, and domain information into $\boldsymbol{U}_m$. To mitigate this, we further apply dimension-wise decorrelation to the joint representation $\boldsymbol{Z}_m$. Specifically, we compute the correlation matrix $\boldsymbol{C}_m \in \mathbb{R}^{p \times p}$ using all representations $\{\boldsymbol{z}_{m,n}^{d_{m,n}}\}_{n=1}^{N_S+N_F}$. Details of the computation for $\boldsymbol{C}_m$ are provided in the appendix. By penalizing the off-diagonal entries of $\boldsymbol{C}_m$, we encourage statistical independence across dimensions, leading to the following decorrelation loss:

$$\mathcal{L}^{\text{DEC}}(\boldsymbol{\theta}) := \sum_{m=1}^{M} ||\boldsymbol{C}_m - \text{diag}(\boldsymbol{C}_m)||_F^2. \tag{6}$$

C. Domain Augmentation. We maintain a domain representation bank $\mathcal{V}_m$ for each modality m, which stores domain representation $\boldsymbol{V}_m$ collected during training, as illustrated in Figure 3. At each training step, newly obtained domain representations are added to the bank, while the oldest ones are discarded to maintain a fixed length. Simultaneously, we sample representations from the banks according to target domains in $\mathcal{D}_F \cup \mathcal{D}_U$. For instance, to emulate a sample from a U-target domain $\boldsymbol{d} = [1, 2]$, we sample domain representation from bank $\mathcal{V}_1$ and $\mathcal{V}_2$ and concatenate them with task-relevant representation of a source domain sample. Training the multimodal classifier

$q(\boldsymbol{Y}|\boldsymbol{Z}_{M+1})$ with these augmented samples exposes it to all possible multimodal domains, thus enhancing the robustness and generalization of the classifier. Let $\hat{\boldsymbol{Y}}'_{M+1}$ denote predicted labels for these augmented samples. The corresponding augmentation loss is defined as

$$\mathcal{L}^{\text{AUG}}_{M+1}(\boldsymbol{\theta}) := -\frac{1}{N_S}\sum_{n=1}^{N_S}\sum_{i=1}^{C}[\boldsymbol{y}_n]_i \log[\hat{\boldsymbol{y}}'_{M+1,n}]_i. \quad (7)$$

The number of augmented samples for each target domain is $N_S/|\mathcal{D}_U \cup \mathcal{D}_F|$, which can vary in practice.

The loss function for the final multimodal classifier trained on original and augmented samples is as following:

$$\mathcal{L}^{\text{CLS}}(\boldsymbol{\theta}) = \mathcal{L}^{\text{TAS}}_{M+1}(\boldsymbol{\theta}) + \mathcal{L}^{\text{AUG}}_{M+1}(\boldsymbol{\theta}). \quad (8)$$

D. Pseudo-labeling. For unlabeled samples from F-target domains, only one modality undergoes a domain shift, while the others remain in the source domain. This observation allows us to use the unshifted modalities to predict labels and treat the resulting predictions as pseudo labels for training. To be specific, for any sample $n \in \{N_S+1, N_S+2, \cdots, N_S+N_F\}$, suppose modality m' is the one undergoing a domain shift. We use the predictions from the remaining modalities $\mathcal{M}' = \{m \in \mathcal{M} \mid m \neq m'\}$ to construct a pseudo label by averaging their outputs:

$$\bar{\boldsymbol{y}}_n = \frac{1}{M-1}\sum_{m\in\mathcal{M}'}\hat{\boldsymbol{y}}_{m,n}. \quad (9)$$

As pseudo labels with lower entropy are typically more reliable, we compute the sample-wise weight $\boldsymbol{w} = [w_{N_S+1}, \cdots, w_{N_S+N_F}]$ using entropy-based softmax normalization:

$$\boldsymbol{w} = \text{softmax}(-[e_{N_S+1}, \cdots, e_{N_S+N_F}]), \quad (10)$$

where e_n denotes entropy of the pseudo label $\boldsymbol{y}'_n, n \in \{N_S+1, \cdots, N_S+N_F\}$.

These pseudo labels are subsequently used to supervise the predictions of both the shifted modality $\hat{\boldsymbol{y}}_{m',n}$ and the aggregated multimodal classifier $\hat{\boldsymbol{y}}_{M+1,n}$, resulting in the following weighted cross-entropy loss:

$$\mathcal{L}^{\text{PLS}}(\boldsymbol{\theta}) = -\sum_{n=N_S}^{N_S+N_F}\sum_{i=1}^{C} w_n[\bar{\boldsymbol{y}}_n]_i \left(\log[\hat{\boldsymbol{y}}_{m',n}]_i + \log[\hat{\boldsymbol{y}}_{M+1,n}]_i\right). \quad (11)$$

The overall loss function is the sum of above losses:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^{\text{CLS}}(\boldsymbol{\theta}) + \alpha_1\mathcal{L}^{\text{SEP}}(\boldsymbol{\theta}) + \alpha_2\mathcal{L}^{\text{DEC}}(\boldsymbol{\theta}) + \alpha_3\mathcal{L}^{\text{PLS}}(\boldsymbol{\theta}), \quad (12)$$

where α_1, α_2 and α_3 are constant coefficients balancing the loss terms.

4 Numerical Results

4.1 Benchmark datasets and baseline models

Benchmark datasets: We evaluate our method on two benchmark datasets: IEMOCAP (Busso et al., 2008), which contains acoustic, visual, and lexical modalities; and KINETICS50-C (Carreira and Zisserman, 2017; Yang et al., 2024), which includes acoustic and visual modalities. IEMOCAP is for the emotion recognition task, composed of scripted and spontaneous dyadic conversations between actors. Following work (Zhao et al., 2021), we select samples from the four classes — neutral, happy, sad and angry, to construct the dataset for our experiments. KINETICS contains 400 human action classes, with at least 400 video clips for each action. KINETICS50-C is a curated subset comprising 50 selected categories from the full dataset.

We follow the data corruption strategies in work (Hendrycks and Dietterich, 2019) to construct target domain samples. Specifically, in work (Hendrycks and Dietterich, 2019), six corruption types are used for the acoustic modality and fifteen for the visual modality, including examples such as windy, crowd noise, and Gaussian noise (acoustic), and fog, rain, and motion blur (visual). For the lexical modality in the IEMOCAP dataset, we randomly mask 20% of the words in each sentence to simulate speech-to-text failure. The source domain samples remained uncorrupted. In IEMOCAP, we select two types of corruption for both the acoustic and visual modalities, and one for the lexical modality, yielding $3 \times 3 \times 2 = 18$ multimodal domains: 1 source domain, 5 F-target domains, and 12 U-target domains. For the KINETICS-C50 dataset, two corruption types are used for both acoustic and visual modalities, resulting in $3 \times 3 = 9$ multimodal domains:1 source, 4 F-target, and 4 U-target domains.

Baseline models: We compare our model, SEAL, with DANN (Ganin et al., 2016), CDAN (Long et al., 2018), ITA (Gholami et al., 2020), CCL (Isobe et al., 2021), DALN (Chen et al., 2022a), DADA(Ren et al., 2024), PCL (Li et al., 2024b), f-DD (Wang and Mao, 2024) and A^3D^2 (Sun et al.,

Method	IEMOCAP				KINETICS50-C			
	Source	Avg(a)	Avg(F)	Avg(U)	Source	Avg(a)	Avg(F)	Avg(U)
DT	75.81	49.23	68.48	41.21	72.60	54.38	61.28	47.48
DANN	77.66	54.02	71.60	46.69	75.07	59.35	64.72	53.98
CDAN	75.02	57.61	70.21	52.36	75.23	59.53	65.40	53.66
ITA	76.71	53.01	72.50	44.89	**76.21**	59.33	66.36	52.30
CCL	**78.73**	57.28	72.49	50.94	70.84	55.15	60.50	49.80
DALN	75.93	57.32	71.70	51.32	26.45	21.50	23.66	19.34
DADA	76.61	54.40	69.50	48.12	75.14	59.12	65.80	52.44
PCL	76.75	51.93	69.52	44.60	75.53	60.61	66.98	54.23
f-DD	74.21	55.94	72.19	49.17	74.87	59.06	65.30	52.82
A^3D^2	77.30	54.21	72.07	46.77	70.14	55.42	60.48	50.36
SEAL	77.12	**60.75**	**72.67**	**55.78**	75.72	**62.51**	**67.56**	**57.47**

Table 1: Comparisons with baseline methods in terms of F1 score (the highest and second-highest scores in each column are highlighted in bold and in blue color, respectively)

Method	F-target domains					U-target domains											
	100	200	010	020	001	011	021	101	110	120	201	210	220	111	121	211	221
DT	56.34	65.19	72.45	75.91	72.52	47.09	60.81	46.65	25.60	62.26	54.69	40.01	48.37	19.28	32.69	19.45	37.63
DANN	69.24	65.75	76.30	78.93	67.78	49.25	56.55	53.35	49.08	74.42	50.30	47.00	48.09	26.43	38.13	24.07	43.64
CDAN	68.69	64.62	70.04	72.20	75.50	52.77	57.61	54.45	42.19	74.19	59.33	51.54	60.26	41.07	45.37	**40.76**	48.81
ITA	69.20	69.56	70.99	79.63	73.15	41.13	55.98	52.66	29.80	64.17	57.03	48.21	49.20	19.98	40.71	33.18	46.67
CCL	68.20	61.80	**82.48**	78.21	71.77	53.45	67.93	42.51	63.24	77.88	42.94	54.27	58.59	22.06	38.57	36.68	53.22
DALN	**73.99**	67.41	77.14	76.69	63.27	51.97	62.42	**63.49**	47.11	66.79	46.97	48.64	51.20	44.11	**48.42**	40.54	44.23
DADA	65.39	64.07	65.43	76.69	75.91	50.59	**68.84**	52.29	49.93	54.40	56.47	37.74	56.33	35.41	37.43	29.92	48.05
PCL	68.70	57.94	73.70	**80.08**	67.17	42.47	57.28	51.03	45.46	**78.05**	50.48	31.45	50.08	30.43	47.82	15.88	34.80
f-DD	68.97	64.81	71.82	76.76	**78.58**	45.20	56.30	60.68	50.28	70.85	59.46	50.89	44.89	31.78	42.80	37.13	39.80
A^3D^2	69.08	69.04	71.62	79.73	70.89	55.52	55.55	48.87	21.07	68.03	63.24	51.28	57.19	18.76	34.90	39.78	47.05
SEAL	64.93	**72.10**	78.67	74.10	73.50	**59.62**	55.08	59.31	**65.45**	65.54	**64.98**	**55.41**	**62.42**	**46.38**	48.39	37.92	48.84

Table 2: Detailed F1 score on each target domain for the IEMOCAP dataset (e.g., column 100 denotes domain $\boldsymbol{d} = [1, 0, 0]$)

2025), which are introduced in the Related Works section.

4.2 Implementation details

For the IEMOCAP dataset, WavLM (Chen et al., 2022b) followed by a TextCNN is employed as the acoustic feature encoder. For the visual modality, APViT pretrained on the RAF-DB (Li et al., 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. For the KINETICS dataset, the acoustic and visual encoders in CAV-MAE pretained on large scale web data are employed. The parameters in the last three layers of the pretrained models are set to be trainable, with all 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 $\alpha_1 = 0.5, \alpha_2 = 5, \alpha_3 = 0.2$. More details of the implementation can be found from the code in the supplementary material. We use the weighted F1 score as model performance metric, which is obtained by averaging the results from four repeated experiments, conducted on four Nvidia A40 GPUs with memory of 48GB.

4.3 Comparison studies

The F1 score comparisons are reported in Table 1, where DT refers to direct transfer, meaning the model is trained with only the source domain data and tested directly without any domain adaptation technique. The “Source", “Avg(a)", “Avg(F)", and “Avg(U)" columns correspond to the performance on source domain, the average of all target domains, the average of all F-target domains, and the average of all U-target domains, respectively. On the IEMOCAP dataset, it is observed that most baseline methods outperform DT even on the source domain. This is attributed to the inclusion of unlabeled F-target samples during training, which help enhance generalization despite lacking ground truth annotations.

The objective of domain adaptation lies in improving performance on the target domains. In this regard, SEAL demonstrates consistent superiority over all compared methods. Specifically, SEAL achieves an average F1 score of 60.75 across

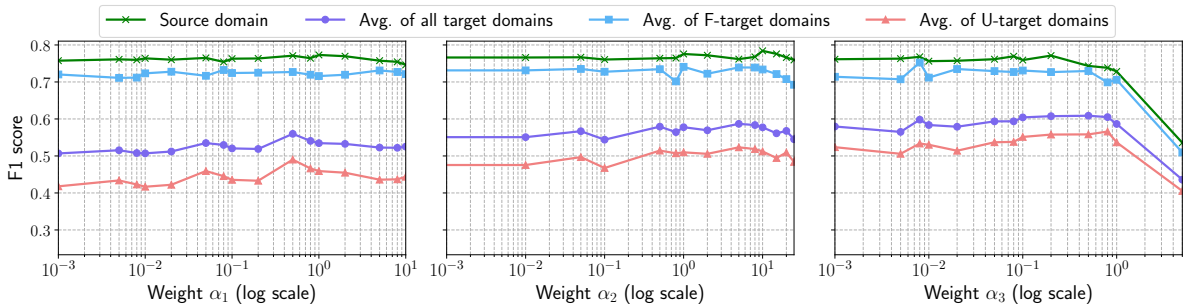

Figure 4: Sensitivity analysis for weights α_1, α_2 and α_3 on the IEMOCAP dataset

all target domains, exceeding the second-highest score (57.61) by more than 3 percentage points. On the F-target domains, SEAL can slightly outperforms the strongest multi-target domain adaptation method, ITA. More notably, on the completely unseen U-target domains, SEAL achieves a substantial improvement of over 3.4 percentage points, validating the effectiveness of our strategy in simulating labeled U-target samples. Similar results are observed on the KINETICS-C50 dataset, corroborating the robustness and generalizability of the proposed approach.

Table 2 exhibits the detailed F1 score on each of the 12 target domains for the IEMOCAP dataset (the results for KINETICS50-C are provided in the appendix). It is shown that SEAL achieves the highest performance on 7 domains and ranks second on 3 others. In comparison, the second best method by this metric, DALN, ranks first on 3 domains and second on 2. These results further highlight the superior overall performance of SEAL across diverse domain shifts.

4.4 Ablation studies

We conduct ablation studies on the IEMOCAP dataset by progressively introducing the loss components $\mathcal{L}^{\text{SEP}}(\boldsymbol{\theta})$, $\mathcal{L}^{\text{DEC}}(\boldsymbol{\theta})$, and $\mathcal{L}^{\text{PLS}}(\boldsymbol{\theta})$. As shown in Table 3, the addition of $\mathcal{L}^{\text{SEP}}$ validates the effectiveness of our separating and augmenting strategy, which exposes the model to simulated unseen U-target domains and boosts the average F1 score from 41.21 to 49.01. Incorporating the disentangling loss $\mathcal{L}^{\text{DEC}}$ further improves performance on both F-target and U-target domains, albeit with a slight decrease on the source domain. Finally, introducing the pseudo-labeling loss $\mathcal{L}^{\text{PLS}}$ benefits both the source and U-target domains. While it results in a minor drop on F-target domains, likely due to label noise, it yields a significant gain on U-target domains, ultimately increasing the average F1 score across all target domains from 58.68 to 60.75.

Loss terms	IEMOCAP			
	Source	Avg(a)	Avg(F)	Avg(U)
$\mathcal{L}^{\text{CLS}}$	75.81	49.23	68.48	41.21
$\mathcal{L}^{\text{CLS}} + \mathcal{L}^{\text{SEP}}$	77.10	55.98	72.69	49.01
$\mathcal{L}^{\text{CLS}} + \mathcal{L}^{\text{SEP}} + \mathcal{L}^{\text{DEC}}$	76.18	58.68	73.92	52.33
$\mathcal{L}^{\text{CLS}} + \mathcal{L}^{\text{SEP}} + \mathcal{L}^{\text{DEC}} + \mathcal{L}^{\text{PLS}}$	77.12	60.75	72.67	55.78

Table 3: The ablation study on the IEMOCAP dataset

Figure 4 presents the sensitivity analysis of the loss weights α_1, α_2, and α_3 on the IEMOCAP dataset. To cover a broad range of values, the x-axis is shown on a logarithmic scale. For all three weights, we observe a similar trend: the F1 scores for different domains initially increase and then decrease, which illustrates how each loss term influences the model performance. The results also demonstrate that our method, SEAL, exhibits robustness across a wide range of hyperparameter values. The optimal configuration, $\alpha_1 = 0.5$, $\alpha_2 = 5$, and $\alpha_3 = 0.2$, can be identified from the peak performance regions in Figure 4.

5 Conclusions

This paper investigates the problem of multimodal multi-target domain adaptation, with the goal of developing an efficient approach that requires only samples from F-target domains, thereby achieving linear sample complexity. To this end, we propose a novel "separate-and-augment with pseudo-labeling" strategy. Specifically, we decompose the representation of each modality into task-relevant and domain-relevant components, and simulate labeled target domain samples by augmenting source-domain task-relevant features with target-domain domain-relevant ones. In addition, we leverage the unshifted modalities in F-target domains to generate pseudo labels for training. Extensive experiments on widely used benchmark datasets demonstrate the effectiveness and robustness of the proposed method.

Limitations

The limitations of this paper are mainly twofold:
1) No theoretical generalization bound for the adaptation method is established.
2) The negative impact of erroneous pseudo-labeling is not considered in our work. In order to focus on our task, we do not have an in-depth investigation on this issue; nevertheless, existing approaches that improve the pseudo-labeling can be incorporated into our framework.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. 2019. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3296–3303.

Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. 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*, pages 7181–7190.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 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.

Xuesong Chen, Shaoshuai Shi, Tao Ma, Jingqiu Zhou, Simon See, Ka Chun Cheung, and Hongsheng Li. 2025. M3net: Multimodal multi-task learning for 3d detection, segmentation, and occupancy prediction in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2275–2283.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. 2025. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *arXiv preprint arXiv:2501.18592*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. 2020. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. 2025. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3779–3787.

Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. 2021. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8187–8196.

Jing Jiang, Sicheng Zhao, Jiankun Zhu, Wenbo Tang, Zhaopan Xu, Jidong Yang, Guoping Liu, Tengfei Xing, Pengfei Xu, and Hongxun Yao. 2025. Multi-source domain adaptation for panoramic semantic segmentation. *Information Fusion*, 117:102909.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. 2024a. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01):1–22.

Junjie Li, Yixin Zhang, Zilei Wang, Saihui Hou, Keyu Tu, and Man Zhang. 2024b. Probabilistic contrastive learning for domain adaptation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 1001–1009.

Shan Li, Weihong Deng, and JunPing Du. 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*, pages 2852–2861.

Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. 2020. Multi-source domain adaptation for visual sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2661–2668.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Jiangbo Pei, Aidong Men, Yang Liu, Xiahai Zhuang, and Qingchao Chen. 2024. Evidential multi-source-free unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5288–5305.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International conference on machine learning*, pages 5171–5180. PMLR.

Chuan-Xian Ren, Yong-Hui Liu, Xi-Wen Zhang, and Ke-Kun Huang. 2022. Multi-source unsupervised domain adaptation via pseudo target domain. *IEEE Transactions on Image Processing*, 31:2122–2135.

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

Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. 2021. Curriculum graph co-teaching for multi-target domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5351–5360.

Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. 2021. Multi-target adversarial frameworks for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9072–9081.

Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, and Kun Zhang. 2021. Domain adaptation with invariant representation learning: What transformations to learn? *Advances in Neural Information Processing Systems*, 34:24791–24803.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Baochen Sun and Kate Saenko. 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*, pages 443–450. Springer.

Jun Sun, Xinxin Zhang, Simin Hong, Jian Zhu, and Lingfang Zeng. 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)*, pages 19680–19690.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Pei Wang, Yun Yang, Yuelong Xia, Kun Wang, Xingyi Zhang, and Song Wang. 2023. Information maximizing adaptation network with label distribution priors for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 25:6026–6039.

Ziqiao Wang and Yongyi Mao. 2024. On f-divergence principled domain adaptation: An improved framework. *arXiv preprint arXiv:2402.01887*.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021. Conditional adversarial networks for multi-domain text classification. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 16–27.

Nishant Yadav, Mahbubul Alam, Ahmed Farahat, Dipanjan Ghosh, Chetan Gupta, and Auroop R Ganguly. 2023. Cda: Contrastive-adversarial domain adaptation. *arXiv preprint arXiv:2301.03826*.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. 2024. Test-time adaptation against multi-modal reliability bias. In *The twelfth international conference on learning representations*.

Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. 2022. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1992–2003.

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

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*.

Xinyu Zhang, Meng Kang, and Shuai Lü. 2024. Low category uncertainty and high training potential instance learning for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16881–16889.

Jinming Zhao, Ruichen Li, and Qin Jin. 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)*, pages 2608–2618.

Sicheng Zhao, Jing Jiang, Wenbo Tang, Jiankun Zhu, Hui Chen, Pengfei Xu, Björn W Schuller, Jianhua Tao, Hongxun Yao, and Guiguang Ding. 2025. Multi-source multi-modal domain adaptation. *Information Fusion*, 117:102862.

Jinjing Zhu, Haotian Bai, and Lin Wang. 2023. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3561–3571.

A Appendix

A1. Table of notations

To improve readability, Table 4 (on the next page) lists the important notations used in the paper along with their definitions, descriptions, values, or ranges.

A2. The calculation of correlation matrix C_m

Let $\hat{\boldsymbol{Z}}$ be the feature matrix that concatenates all feature $\{\boldsymbol{z}_{m,n}^{d_{m,n}}\}_{n=1}^{N_S+N_F}$, that is,

$$\hat{\mathbf{Z}}_m := \begin{bmatrix} \left(\boldsymbol{z}_{m,1}^{d_{m,1}}\right)^\top \\ \left(\boldsymbol{z}_{m,2}^{d_{m,2}}\right)^\top \\ \vdots \\ \left(\boldsymbol{z}_{m,N_S+N_F}^{d_{m,N_S+N_F}}\right)^\top \end{bmatrix} \in \mathbb{R}^{(N_S+N_F)\times p}.$$

We first center the features by subtracting the mean of each column:

$$\bar{\mathbf{Z}}_m = \hat{\mathbf{Z}}_m - \mathbf{1}\boldsymbol{\mu}_m^\top,$$

where $\boldsymbol{\mu}_m = \frac{1}{N_S+N_F}\sum_{n=1}^{N_S+N_F} \boldsymbol{z}_{m,n}^{d_{m,n}}$.

The empirical covariance matrix is computed as:

$$\bar{\mathbf{C}}_m = \frac{1}{N_S+N_F-1}\bar{\mathbf{Z}}_m^\top \bar{\mathbf{Z}}_m.$$

Let $\boldsymbol{\sigma}_m \in \mathbb{R}^p$ be the vector of standard deviations of each feature:

$$\boldsymbol{\sigma}_m = \sqrt{\mathrm{diag}(\bar{\mathbf{C}}_m)}.$$

Then, the correlation matrix is given by:

$$\mathbf{C}_m = \mathbf{D}_m^{-1}\bar{\mathbf{C}}_m\mathbf{D}_m^{-1}, \quad \text{where } \mathbf{D}_m = \mathrm{diag}(\boldsymbol{\sigma}_m).$$

A3. Detailed comparisons on all target domains for the KINETICS50-C dataset

Table 5 reports the detailed F1 score on each of the 8 target domains for the KINETICS50-C dataset. SEAL ranks first on 5 domains and second on 1 domain. In comparison, the second-best method, PCL, ranks first on 2 domains and second on 1. These results again demonstrate the superior performance of SEAL.

Notation	Definition / Description / Value / Range
m	modality index; $m \in \{1, 2, \cdots, M\}$
$\boldsymbol{x}_m$	the raw feature of modality m
d_m	unimodal domain index (label); $d_m \in \{0, 1, \cdots, D_m - 1\}$ or corresponding one-hot vector
$\boldsymbol{d}$	multimodal domain indicator; $\boldsymbol{d} := [d_1, d_2, \cdots, d_M]$
$\mathcal{D}_S$	the set of source domain indicator; $\mathcal{D}_S := \{\boldsymbol{d} \vert d_1 = d_2 = \cdots = d_M = 0\}$
$\mathcal{D}_F$	the set of F-target domain indicators; $\mathcal{D}_F := \{\boldsymbol{d} \vert \sum_{m=1}^{M} \mathbf{1}_{(d_m \neq 0)} = 1\}$
$\mathcal{D}_U$	the set of U-target domain indicators; $\mathcal{D}_U := \{\boldsymbol{d} \vert \sum_{m=1}^{M} \mathbf{1}_{(d_m \neq 0)} \geq 2\}$
n	sample index; $n \in \{1, 2, \cdots, N_S + N_F\}$
$d_{m,n}$	domain index (label) of modality m; $d_{m,n} \in \{0, 1, \cdots, D_m - 1\}$ or corresponding one-hot vector
$\boldsymbol{d}_n$	domain indicator of sample n; $\boldsymbol{d}_n = [d_{1,n}, d_{2,n}, \cdots, d_{M,n}]$
$\boldsymbol{x}_{m,n}^{d_{m,n}}$	the raw feature of the n-th sample
$\boldsymbol{x}_n^{\boldsymbol{d}_n}$	all raw feature of the n-th sample; $\boldsymbol{x}_n^{\boldsymbol{d}_n} = \{\boldsymbol{x}_{m,n}^{d_{m,n}}\}_{m=1}^{M}$
$\boldsymbol{y}_n$	task label of the n-th sample; scalar value of class or corresponding one-hot vector
$\boldsymbol{z}_{m,n}^{d_{m,n}}$	representation derived from the raw feature $\boldsymbol{x}_{m,n}^{d_{m,n}}$
$\boldsymbol{u}_{m,n}^{d_{m,n}}$	task-relevant representation contained in $\boldsymbol{z}_{m,n}^{d_{m,n}}$
$\boldsymbol{v}_{m,n}^{d_{m,n}}$	domain-relevant representation contained in $\boldsymbol{z}_{m,n}^{d_{m,n}}$
$\boldsymbol{X}_m^{d_m}$	random variable of the raw feature, with $\boldsymbol{x}_{m,n}^{d_{m,n}}$ as its instance
$\boldsymbol{Y}$	random variable of the task label, with $\boldsymbol{y}_n$ as its instance
$\boldsymbol{D}_m$	random variable of the domain label, with $d_{m,n}$ as its instance
$\boldsymbol{Z}_m^{d_m}$	random variable of the representation, with $\boldsymbol{z}_{m,n}^{d_{m,n}}$ as its instance
$\boldsymbol{U}_m^{d_m}$	random variable of the task-relevant representation, with $\boldsymbol{u}_{m,n}^{d_{m,n}}$ as its instance
$\boldsymbol{V}_m^{d_m}$	random variable of the domain-relevant representation, with $\boldsymbol{v}_{m,n}^{d_{m,n}}$ as its instance

Table 4: Notations and their definitions/descriptions/values/ranges

Method	F-target domains				U-target domains			
	10	20	01	02	11	12	21	22
DT	65.20	66.35	55.58	58.01	44.70	53.77	42.63	48.81
DANN	67.70	69.42	59.70	62.08	53.60	60.54	46.08	55.7
CDAN	67.60	69.41	60.00	64.60	50.02	59.31	48.36	56.95
ITA	68.70	70.08	62.30	64.36	50.77	61.59	40.54	56.31
CCL	65.43	66.40	50.67	59.51	44.83	58.87	42.24	53.25
DALN	23.83	20.65	24.79	25.36	21.36	21.22	16.20	18.59
DADA	69.06	71.00	58.77	64.38	50.03	59.46	45.95	54.33
PCL	**70.22**	**73.79**	61.28	62.64	50.72	60.28	49.71	56.19
f-DD	67.86	70.14	58.59	64.60	49.69	61.01	42.83	**57.74**
A^3D^2	63.20	65.53	54.55	58.64	47.76	54.76	45.88	53.03
SEAL	68.63	71.64	**65.26**	**64.69**	**57.64**	**62.41**	**54.44**	55.37

Table 5: Detailed F1 score on each target domain for the KINETICS50-C dataset (the highest and second-highest scores in each column are highlighted in bold and in blue color, respectively)