Domain Adaptation under Continuous Spurious Shift

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in domain adaptation have shown promise in transferring knowledge across domains characterized by a continuous value or vector, such as varying patient ages, where "age" serves as a continuous index. However, these approaches often fail when spurious features shift continuously along with the domain index. This paper introduces a new method designed to withstand the continuous shifting of spurious features during domain adaptation. Our method enhances domain adaptation performance by aligning representations across continuously indexed domains, inspired by principles of causal transportability. Theoretical analysis provides insight into how our approach encourages transportable representations across different domains under certain assumptions. Empirical results, from both semi-synthetic and real-world medical datasets, indicate that our method outperforms state-of-the-art domain adaptation methods.

1 Introduction

Machine learning typically presumes that the training and test data come from identical distributions, hoping that the trained model effectively generalizes to the test environment. However, this presumption breaks when the training and testing occur across different domains (e.g., with distinct source and target domains). Domain adaptation (DA) effectively addresses this challenge by utilizing labeled data from the source domain along with either unlabeled or minimally labeled data from the target domain, thereby improving model performance (Ben-David et al., 2010; Ganin et al., 2016; Tzeng et al., 2017; Zhang et al., 2019).

Continuously indexed domain adaptation (CIDA) (Wang et al., 2020) generalizes typical DA, which focuses on discrete domains (transferring from dataset A to B), to DA across continuously indexed domains, where domain shift is characterized by a continuous index such as time and location. For instance, in healthcare, CIDA can be instrumental in transferring knowledge across patient data that varies with age. As patients age, their physiological parameters and response to treatments can change subtly; in response, CIDA aims to train a model that can adapt to these continuous shifts. Unfortunately, previous DA methods (Ganin et al., 2016; Tzeng et al., 2017; Zhang et al., 2019), including CIDA (Wang et al., 2020), often fail when spurious features are continuously shifting.

Example 1 (**Continuously Shifting Spurious Features in Sleep Studies**). *Suppose one trains a model to take as input a time series of breathing signal* $\mathbf{x} \in \mathbb{R}^T$ *to predict the corresponding sleep stage as* $y^{pred} \in \{$*'Awake'*, *'Light Sleep'*, *'Deep Sleep'*, *'REM Sleep'*$\}$. *Here 'respiratory rate' is a typical spurious feature. If one trains the model using data from young subjects (with age* $20 \sim 30$*), it will learn that 'higher respiratory rates' often correspond to 'REM Sleep'. However, as subjects continuously age, they may have a slower respiratory rate in breathing patterns. Therefore the spurious feature 'respiratory rate' no longer works when predicting sleep stages for older subjects, and the older the subject is, the lower the model accuracy.*

Our analysis suggests that neither alignment-based methods (Ganin et al., 2016; Wang et al., 2020) nor causality-inspired methods (Mao et al., 2022) alone fully address this problem. Alignment-based methods tend to align spurious features rather than underlying causal features, and may therefore fail to generalize across continuously indexed domains with continuously shifting spurious features. On the other hand, causality-inspired methods aiming to learn causal features (encodings) may in practice behave similarly to association-based methods (see Sec. 3 for detailed analysis and Sec. 6 for supporting empirical results), and

Table 1: Comparison of our CADA with different representative previous methods.

	CUA	ADDA	DANN	CDANN	MDD	CIDA	VOOD	GDA	CADA (Ours)
Continuous	✗	✗	✗	✗	✗	✓	✗	✓	✓
Causal	✗	✗	✗	✗	✗	✗	✓	✗	✓
Multi-Domain	✓	✗	✓	✓	✗	✓	✗	✓	✓
Covariate Shift	✓	✓	✓	✓	✓	✓	✗	✓	✓
Setting (DA/DG)	DA	DA	DA	DA	DA	DA	DG	DA	DA

therefore may also fail to generalize. Motivated by such analysis, we propose to jointly (1) learn representations inspired by causal transportability and (2) align these representations across continuously indexed domains, thereby improving domain adaptation performance under continuous spurious shift. Our contributions are as follows:

- We identify the problem of continuous spurious shift and propose Continuously trAnsportable Domain Adaptation (CADA) as the first general, causality-inspired DA method to address this problem.
- Our theoretical analysis provides insight into how CADA encourages transportable representations across continuously indexed domains under certain assumptions.
- Empirical results on both semi-synthetic and real-world medical datasets show that our method outperforms the state-of-the-art DA methods in the face of continuous spurious shift.

2 Related Work

Typical Domain Adaptation. Domain adaptation has long been studied (Farahani et al., 2021; Csurka, 2017; Ben-David et al., 2010; Peng et al., 2019; Prabhu et al., 2021; Liu et al., 2023; Xu et al., 2022) to promote model's generalization ability on unseen domains, with unlabeled or limited labeled data. Traditional methods include importance weighting (Shimodaira, 2000; Gretton et al., 2009; Lipton et al., 2018), self-training (Zou et al., 2018; Kumar et al., 2020; Prabhu et al., 2021), distribution matching (Pan et al., 2010; Tzeng et al., 2014; Sun & Saenko, 2016; Peng et al., 2019; Nguyen-Meidine et al., 2021; He et al., 2024) and adversarial-based training (Ganin et al., 2016; Zhang et al., 2019; Zhao et al., 2017; Wang et al., 2020; Xu et al., 2023). Typically, these methods focus on categorical domains, where a one-hot vector indicates which domain a data point comes from. In contrast, CIDA (Wang et al., 2020) extends to continuous domains, with a continuous variable specifying each data point's domain. CIDA seeks to align encodings of different domains based on this continuous identification. However, during the alignment, CIDA may struggle to eliminate the influence of continuously shifting spurious features, leading to poor performance (see empirical results in Sec. 6). In contrast, our method mitigates the impact of such spurious features by learning and aligning representations inspired by causal transportability.

Causality-Inspired Domain Adaptation. Causal inference is a powerful method for modeling knowledge transfer (Bareinboim & Pearl, 2014; 2013; 2016; Bühlmann, 2020; Correa & Bareinboim, 2019; 2020; Magliacane et al., 2018; Rojas-Carulla et al., 2018) and ensuring structure invariance (Pearl et al., 2000). Various studies have explored its application to eliminate spurious features and improve domain adaptation performance (Arjovsky et al., 2019; Mahajan et al., 2021; Mao et al., 2021; Yue et al., 2021). For instance, VOOD (Mao et al., 2022) aims to estimate invariant causal effect across different environments, hoping to improve domain generalization (DG) performance. However, these causal methods (1) are still subject to covariate shift, often compromising the causal models' identifiability (Pearl et al., 2000), (2) can easily degenerate to association-based method, and (3) are not designed for continuously indexed domains. In contrast, our CADA effectively addresses these problems via joint latent-space distribution alignment and causal inference.

Comparison of Representative Methods. Table 1 summarizes the differences between our CADA and representative previous methods in terms of whether they (1) are designed for continuously indexed domains, (2) are causal/robust against spurious features, (3) can naturally handle multiple source/target domains, (4) are robust against covariate shift, and (5) are designed for DA or DG.

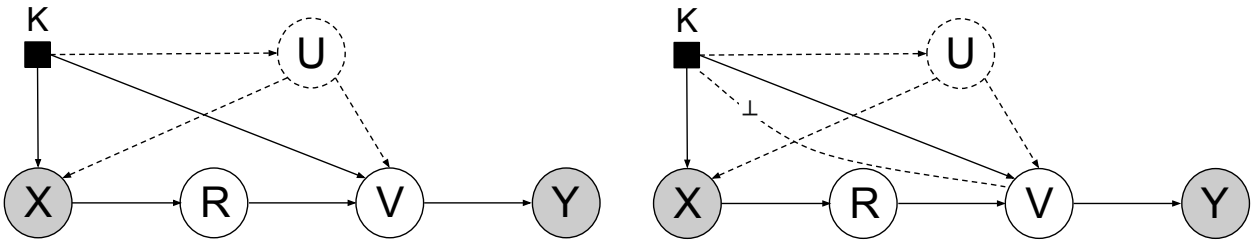


Figure 1: **Left**: Causality-inspired graphical model for CADA (note that the full model includes an additional component that enforces conditional independence $K \perp V$). Shaded and transparent circles denote observed and latent variables, respectively. Dotted circle denotes an unobserved confounder. Dotted arrows denote unobserved dependence from or to the confounder. **Right**: Causality-inspired graphical model for the full CADA model. A dashed edge with a $\perp$ sign between K and V indicates the enforced independence $K \perp V$.

3 Theoretical Analysis

In this section, we describe the challenges of performing domain adaptation on data with continuously shifting spurious correlations and present a causality-inspired perspective on the problem. We then present the solution in Sec. 4.

Notation and Problem. We consider the problem of unsupervised continuously indexed domain adaptation as proposed in (Wang et al., 2020). We assume that the continuous domain index set $\mathcal{K}$ is a part of a metric space, $\mathcal{K} = \mathcal{K}_s \cup \mathcal{K}_t$, with $\mathcal{K}_s$ and $\mathcal{K}_t$ as domain index sets for the source and target domains, respectively. The inputs and labels are denoted as $\mathbf{x}$ and y, respectively. Given the labeled source-domain data $(\mathbf{x}_i^s, y_i^s, k_i^s)_{i=1}^n$ and unlabeled target-domain data $(\mathbf{x}_i^t, k_i^t)_{i=1}^l$, with $k_i^s \in \mathcal{K}_s$ and $k_i^t \in \mathcal{K}_t$, the goal is to predict the target-domain labels $(y_i^t)_{i=1}^l$. (Note that the labels $(y_i^t)_{i=1}^l$ are only available at test time for evaluation purpose only.) We use upper-case letters (e.g., X) to denote random variables and lower-case letters (e.g., x) to denote corresponding realizations.

3.1 A Causality-Inspired View on Continuously Indexed Domain Adaptation

Structural Causal Model. We use a graphical model similar to a Structural Causal Model (SCM) in Fig. 1 to describe the underlying generative process of the random variables of the data X and their labels Y[1]. In addition to the observed variables X and Y, this SCM also involves the following variables:

- R, which is a vector encoding both causal and spurious features of the input data X,
- V, which is a representation derived from R and may contain both causal and spurious components prior to intervention,
- U, which is the unobserved noise and spurious features from external sources, and
- K, which is the continuous domain index.

U and K together affect both the generation of input X and the generation of V from R. As a result, V may still contain spurious components induced by U and K. The interventional distribution $P(V \mid \mathrm{do}(X))$ motivates a representation that aims to reduce these spurious influences and serves as the basis of our prediction model. However, using this interventional quantity alone is not sufficient in practice (more details below).

Note that in the context of causal inference, K can be considered as a continuous version of the *selection node* (Pearl & Bareinboim, 2011); it points to variables whose generation process differs for different domains. In particular, in the SCM of Fig. 1, we treat $P(X)$ and $P(V \mid K, R)$ as domain-specific mechanisms, while $P(R \mid X)$ and $P(Y \mid V)$ are assumed to be invariant across domains.

Moreover, R and V denote latent semantic variables in the underlying data-generating process. The encoders in our model are designed to learn representations that approximately follow the structural relationships specified in the SCM.

[1]Note that our CADA is a *causality-inspired* model rather than a typical *causal* model.

It is also worth noting that the interventional distribution of V does *not* automatically eliminate spurious signals. Rather, the goal of our framework is to construct a representation whose interventional distribution removes the spurious correlations induced by the domain variable, through a *joint* formulation of **an interventional distribution** and **a minimax game**, as described in Sec. 4.

Following Example 1, below we provide an example of using the SCM in Fig. 1 for sleep studies across continuously indexed domains (patients of different ages).

Example 2 (**Causal Model for Sleep Studies**). *In sleep studies (Zhao et al., 2017), one typical task is to estimate the sleep stage Y (e.g., "Awake" and "Deep Sleep") from a patient's breathing signal X; here "age" is a domain index K; patients of different ages belong to different domains. One robust way to predict Y is to use breathing patterns such as periodicity (periodic breathing signals usually indicate "Deep Sleep"); these patterns are encoded in the representations V in our CADA. The unobserved **spurious features** U in sleep studies may include 'respiratory rate' mentioned in Example 1. Breathing signals X from older patients, i.e., larger K, may have a lower 'respiratory rate' due to weakened respiratory muscles; therefore it is **not a reliable feature** for predicting Y. A neural network can extract the compact representation R given the raw breathing signal X as input, but R still inherits the spurious features U from X; therefore directly using R to predict Y would not generalize across domains (different K), as in Example 1. To address this problem, we aim to construct a representation V from R and then use the interventional distribution $P(V \mid do(X))$ **with a minimax game (in Sec. 4)** to remove the spurious influence of U and K, thereby obtaining a representation intended to be more transportable across domains for predicting Y.*

3.2 Challenges in Continuously Transportable Domain Adaptation

Conditional Model $P(Y|X)$ Is Not Generalizable across Domains. Due to the confounding factors K and U between X and V, the computable quantity $P(Y|X)$ trained in source domains usually does not generalize to target domains, i.e. $P(Y|X, K = k_1) \neq P(Y|X, K = k_2)$ for $k_1 \neq k_2 \in \mathcal{K}$.

The Interventional Quantity $P(Y|do(X))$ Suggests a Domain-Transportable Objective. Under the SCM in Fig. 1 and its associated assumptions, the interventional quantity $P(Y|do(X))$ is transportable across different domains in principle. This is shown in Theorem 3.1 and Theorem 3.2 below.

Lemma 3.1 (**Front-Door Criterion for R w.r.t. (X, V)**). *Under the SCM in Fig. 1, the variable R satisfies the front-door criterion relative to the ordered pair (X, V).*

Proof. Under the assumed causal graph:

1. R intercepts all directed paths from X to V, since the only directed path is $X \to R \to V$.
2. There is no unblocked back-door path from X to R.
3. All back-door paths from R to V are blocked by X.

Therefore, R satisfies the front-door criterion relative to (X, V). □

Theorem 3.1 (**Transportability of $P(v|do(x))$**). *Under the SCM in Fig. 1, the quantity $P(v \mid do(x))$ is transportable from domain k_1 to k_2 for any $k_1 \neq k_2$.*

Proof. By Lemma 3.1, R satisfies the front-door criterion (Pearl, 2009), we can then use the front-door adjustment formula to obtain:

$$P(v|do(x)) = \sum\nolimits_r P(r|x) \sum\nolimits_{x'} P(v|r, x')P(x').$$

By the definition of trivial transportability (Pearl & Bareinboim, 2011), if $P(v|do(x))$ is identifiable under the assumed SCM, then it is transportable. Specifically, identifiability implies that this interventional quantity can be expressed in terms of observational distributions without explicitly involving the domain index K. Therefore, under these assumptions, $P(v|do(x))$ is invariant across domains. □

Theorem 3.2 (**Transportability of** $P(y|do(x))$)**.** *Under the SCM in Fig. 1, the quantity* $P(y|do(x))$ *is trivially transportable from domain* k_1 *to* k_2 *for any* $k_1 \neq k_2$ *using:*

$$P(y|do(x)) = \sum\nolimits_v P(y|v)Q(v|x), \tag{1}$$

where $Q(v|x)$ *is a probabilistic encoder generating encoding* v *given the input* x*. It is constructed as*

$$Q(v|x) = \sum\nolimits_r P(r|x) \sum\nolimits_{x'} P(v|r, x')P(x'). \tag{2}$$

Proof. Using the chain rule we have that $P(y|do(x)) = \sum_v P(y|v)P(v|do(x))$, where $P(v|do(x))$ is given by Theorem 3.1. Denoting $P(v|do(x))$ as $Q(v|x)$ concludes the proof. □

Transportability Alone Is Not Sufficient. From Eqn. 1, we can see that v computed from Eqn. 2 can be treated as the encoding for the input x. In addition to the assumed causal model, an additional conditional independence $K \perp V$ is required. When this condition is violated, predictions may still depend on the domain index through residual domain-specific variations, which can lead to degraded performance under domain shift. In practice, due to covariate shift (Ganin et al., 2016; Ben-David et al., 2010), there is no guarantee that directly using Eqn. 1 will lead to good domain adaptation performance.

Specifically, while Eqn. 2 motivates a causality-inspired encoder, this interpretation can break down in the presence of covariate shift, which may lead to non-positivity in the distribution of V and undermine the identifiability of the interventional distribution $P(V|do(X))$. This required positivity condition can be stated as follows: for any pair of domain indices k_1 and k_2,

$$p(v \mid K = k_1) > 0 \quad \text{if and only if} \quad p(v \mid K = k_2) > 0,$$

that is, the support of V must be identical across different domains. This matches the standard positivity assumption in Definition 3.2.4 of (Pearl, 2009), which requires that the relevant distributions remain strictly positive across domains. Although the expression

$$P(v \mid do(x)) = \sum\nolimits_r P(r \mid x) \sum\nolimits_{x'} P(v \mid r, x')\, P(x')$$

holds across different domains indexed by K, it is only meaningful when the positivity assumption is satisfied. In practice, enforcing the marginal independence $K \perp V$ ensures identical support of V across domains, since

$$K \perp V \Leftrightarrow p(v \mid K = k_1) = p(v \mid K = k_2), \quad \forall v.$$

Note that positivity/support overlap only requires the support of V to be shared across domains, rather than identical conditional distributions $p(v \mid K = k)$. However, the following lemma shows that reducing the distribution divergence between the encoded representation V across domains can help reduce the generalization error on the target domain. Enforcing independence between K and V can therefore be viewed as a practical way to encourage such alignment across domains.

The following lemma adapted from (Ben-David et al., 2010) shows the consequence of covariate shift.

Lemma 3.2 (**Target-Domain Error Bound**)**.** *Let* $\mathcal{H}$ *be a hypothesis space, and* $h \in \mathcal{H} : \mathcal{V} \to \{0, 1\}$*.* $\mathcal{D}_S(V)$ *and* $\mathcal{D}_T(V)$ *are the encoding distributions of the source and target domains, respectively. We have:*

$$\epsilon_{\mathcal{D}_T}(h) \leq \epsilon_{\mathcal{D}_S}(h) + \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(V), \mathcal{D}_T(V)) + \lambda,$$

where $\epsilon_{\mathcal{D}_S}(h)$ *and* $\epsilon_{\mathcal{D}_T}(h)$ *are the prediction error in the source and target domains, respectively. The constant* $\lambda = \min_h(\epsilon_{\mathcal{D}_S}(h) + \epsilon_{\mathcal{D}_T}(h))$*. The* $\mathcal{H}\Delta\mathcal{H}$ *divergence* $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(V), \mathcal{D}_T(V))$ *characterizes the divergence between the source domain's and the target domain's encoding distributions.*

Lemma 3.2 shows that the generalization error of a target domain is bounded by the source-domain error, the distribution divergence between the encoding v of the source and target domains, and the constant λ. Therefore it is also important to reduce the second term $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S(V), \mathcal{D}_T(V))$ by aligning the encoding distributions of v from different domains. This lemma suggests that reducing the distribution divergence between the encoded representation V across domains can help reduce the generalization error in the target domain.

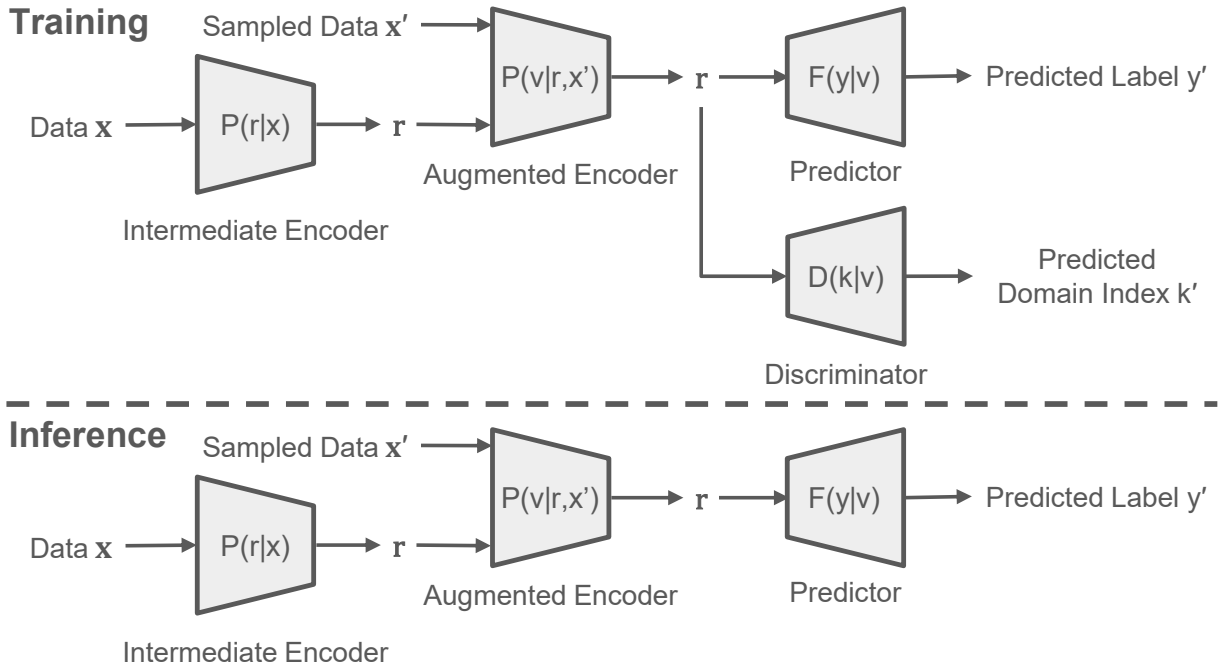


Figure 2: **Schematic overview of CADA.** Given an input x, the causal encoder samples $r \sim P(r|x)$, draws an auxiliary sample $x' \sim P(x')$, and then samples $v \sim P(v|r, x')$. The predictor $F(y|v)$ is trained on labeled source data, while the discriminator $D(k|v)$ is trained adversarially to regress the domain index and encourage domain-invariant representations. At inference time, predictions are obtained by Monte Carlo estimation of $P(y \mid \mathrm{do}(x))$ through the composed encoder.

Definition 1 (**Ideal Encoder** $Q(v|x)$)**.** *With the analysis above, an ideal encoder $Q(v|x)$ that encodes the input x into an encoding v is an encoder that*

(1) has the form of Eqn. 2 and
(2) ensures the marginal independence $k \perp\!\!\!\perp v$.

The first requirement in Definition 1 is trivially satisfied by construction. However, with v generated by $Q(v|x)$ in Eqn. 2, there is no guarantee that $k \perp\!\!\!\perp v$ (the second requirement in Definition 1); this is shown in Theorem 3.3 below.

Theorem 3.3. *With the causal diagram G in Fig. 1, K is not independent with V.*

Proof. This is straightforward given that $K \not\perp\!\!\!\perp V$ in $G_{\overline{X}}$. □

Therefore, to enable transportable domain adaptation, one needs to train an encoder parameterized by Eqn. 2 while aligning the distributions of v for different domains. Such alignment is done via an adversarial training process, as detailed in Sec. 4 below.

Remark. The structural causal model introduced above serves as an abstract representation of latent semantic factors and their relationships. A discussion of the scope and limitations of these modeling assumptions is provided in Appendix A.

4 Method

To learn an ideal encoder $Q(v|x)$ that satisfies the two requirements as defined in Definition 1, we propose to learn a sophisticated encoder $Q(v|x)$ in the form of Eqn. 2 while encouraging the distributions of the representations $v \sim Q(v|x)$ from all domains $\mathcal{K}$ to be aligned. Such alignment corresponds to Requirement (2) of Definition 1, i.e., the marginal independence $k \perp\!\!\!\perp v$; it ensures that all labels can be accurately predicted

by the shared predictor $P(y|v)$ using Eqn. 1. We achieve this using an additional discrimininator $D(k|v)$, which predicts the domain index k given the causal encoding v.

Once an ideal encoder $Q(v|x)$ is trained, one can then use Eqn. 1 to predict y. Below, we start by introducing our CADA's causality-inspired encoder $Q(v|x)$, predictor $F(y|v)$, and discriminator $D(k|v)$. We can put them together to form a final objective function to perform a minimax optimization. Fig. 2 provides a schematic overview of the model composition and the training objective.

Causality-Inspired Encoder $Q(v|x)$. Our causality-inspired encoder will take the form of Eqn. 2, which consists of three components, i.e., an intermediate encoder $P(r|x)$, an augmented encoder $P(v|r,x)$, and a data sampler $P(x')$:

- **Intermediate Encoder $P(r|x)$:** The intermediate encoder encodes the input x into the intermediate representation r. To enable sampling of r, this is a *probabilistic* encoder, where sampling $r \sim P(r|x)$ is equivalent to

$$r \sim \mathcal{N}\big(\mu_r(x), \sigma_r^2(x)\big),$$

  where $\mu_r(\cdot)$ and $\sigma_r^2(\cdot)$ denote two neural networks taking x as input and predict the mean and variance of r, respectively. These networks can be trained using the reparameterization trick (Kingma & Welling, 2013).
- **Augmented Encoder $P(v|r,x')$:** The augmented encoder takes as input the intermediate representation r and another input data point $x' \neq x$ as augmented data to predict the preliminary v. Similar to $P(r|x)$, sampling $v \sim P(v|r,x')$ is equivalent to

$$v \sim \mathcal{N}\big(\mu_v(r,x'), \sigma_v^2(r,x')\big),$$

  where $\mu_v(\cdot,\cdot)$ and $\sigma_v^2(\cdot,\cdot)$ denote two neural networks trained using the reparameterization trick (Kingma & Welling, 2013).
- **Data Sampler $P(x')$:** The data sampling uniformly randomly samples data from different domains without conditioning on labels. Intuitively, the goal is to include diverse data with different spurious features into the model, such that these spurious features can cancel each other out using the summation $\sum_{x'} P(v|r,x')P(x')$ in Eqn. 2, thereby leading to representations v that are less sensitive to domain-specific spurious variation.

As a result, given the input x, sampling v from our causal encoder is equivalent to first sampling x's intermediate representation r from $P(r|x)$, sampling x' from the data sampler $P(x')$, and then sampling v from $P(v|r,x')$. Note that this composed causality-inspired encoder $Q(v|x)$ can be trained end-to-end with the reparameterization trick (Kingma & Welling, 2013).

Predictor $F(y|v)$. According to Eqn. 1, our predictor $F(y|v)$ takes as input the the learned representations v from $Q(v|x)$ and predicts the label y. In CADA, $F(y|v)$ is parameterized by a simple multi-layer perceptron (MLP). Note that $F(y|v) = P(y|v)$ in CADA.

Discriminator $D(k|v)$. Our discriminator $D(k|v)$ takes as input the representation v and predicts the domain index k. $D(k|v)$ is crucial in terms of aligning the distributions of causal encodings v from different domains (more details below). Since we focus on DA across continuously indexed domains, $D(k|v)$ will directly *regress* the continuous index k (Wang et al., 2020) rather than performing classification (Ganin et al., 2016).

Final Objective Function. Putting $Q(v|x)$, $F(y|v)$, and $D(k|v)$ together, we can form the following minimax optimization:

$$\min_{Q,F} \max_{D} V_p(Q,F) - \lambda_d V_d(D,Q), \tag{3}$$

where the instantiations of both terms are **very different** from (Ganin et al., 2016; Wang et al., 2020). Specifically, the **first term**

$$V_p(Q,F) \triangleq \mathbb{E}^s[L_p(\widehat{y}, y)], \tag{4}$$

$$V_p(Q,F) \triangleq \mathbb{E}^s[L_p(\widehat{y},y)] \approx \tfrac{1}{N_s}\sum\nolimits_{(x,y,k)} -\log \sum\nolimits_v F(y|v)Q(v|x), \tag{5}$$

measures the *difference* between the *prediction* $\widehat{y}$ (computed by Q and F) and the *ground truth* y; this can be computed as the negative log-likelihood of y, i.e., $-\log\sum_v F(y|v)Q(v|x)$. We compute the average difference over N_s tuples of (x,y,k) sampled from $p^s(x,y,k)$.

Note that to generate the prediction $\widehat{y}$, one has to go through $Q(v|x)$ and $F(y|v)$, with $Q(v|x)$ consisting of three components $P(r|x)$, $P(x')$, and $P(v|r,x')$ in Eqn. 2. Specifically, given an input-label-index tuple (x,y,k) sampled from the source data distribution, we will first sample r from $P(r|x)$, sample x' from $P(x')$, sample v from $P(v|r,x')$ given the previous sampled r and x', and then generate prediction $\widehat{y}$ from $P(y|v)$ given the sampled v. We then compute the prediction loss for $\widehat{y}$ against the label y.

The **second term**,

$$V_d(D,Q) \triangleq \mathbb{E}[L_d(\widehat{k},k)],$$

$$V_d(D,Q) \triangleq \mathbb{E}[L_d(\widehat{k},k)] \approx \tfrac{1}{N_{all}}\sum\nolimits_{(x,y,k)} -\log \sum\nolimits_v D(k|v)Q(v|x), \tag{6}$$

measures the *difference* between the *predicted* domain index $\widehat{k}$ (computed by Q and D) and the *ground-truth* domain index k; this can be computed as the negative log-likelihood of k, i.e., $-\log\sum_v D(k|v)Q(v|x)$. Similar to Eqn. 5, we compute the average difference over N_{all} data points sampled from $p(x,y,k)$. Specifically, given an input-label-index tuple (x,y,k) sampled from the entire data distribution, we first sample v from $Q(v|x)$ by going through $P(r|x)$, $P(x')$, and $P(v|r,x')$; the discriminator $D(k|v)$ then generates the predicted domain index $\widehat{k}$ given v. See Sec. 5 for more details on the mapping between Eqn. 3 and our causal estimand.

Two Requirements in Definition 1. The minimax optimization with the objective function Eqn. 3 ensures that two requirements in Definition 1 are satisfied:

- Requirement (1) of Definition 1 is ensured by parameterizing the causality-inspired encoder $Q(v|x)$ according to Eqn. 2, which uses two probabilistic neural networks, i.e., the intermediate encoder $P(r|x)$ and the augmented encoder $P(v|r,x')$, and the data sampler $P(x')$.
- Requirement (2) of Definition 1 is encouraged by the adversarial loss, i.e., the minimax optimization; the discriminator $D(k|v)$ will be trained to predict the domain index k, while causality-inspired encoder $Q(v|x)$ is trained to fool the discriminator. This adversarial process will try to align the distributions of v across different domains such that $D(k|v)$ cannot accurately predict the domain index k. As shown in (Wang et al., 2020), the solution where $k \perp\!\!\!\perp v$ will be among the equilibria of the minimax optimization in Eqn. 3.

Inference (Prediction). After our CADA is trained using Eqn. 3, one can then make predictions (causal inference) by combining Eqn. 1 and Eqn. 2. Specifically, given the input x, one can predict y using

$$P(y|do(x)) = \sum_v P(y|v) \sum_r P(r|x) \sum_{x'} P(v|r,x')P(x'),$$

where we use Monte Carlo estimation to draw samples from different conditional distributions and aggregate the results to obtained the final prediction (see Alg. 1 in Appendix B for the detailed algorithm).

5 Discussion

Connection between Eqn. 3 and the Causal Estimand. Our method has two major components jointly trained using the minimax objective function in Eqn. 3, i.e.,

$$\min_{Q,F}\max_{D} V_p(Q,F) - \lambda_d V_d(D,Q). \tag{7}$$

We provide details for these two terms below.

First Term $V_p(Q, F)$. The first term is the negative log-likelihood of predicting the label y, i.e.,

$$V_p(Q, F) = \frac{1}{N_s} \sum_{(x,y,k)} -\log \sum_v F(y \mid v) Q(v \mid x), \tag{8}$$

where

$$Q(v \mid x) = \sum_r P(r \mid x) \sum_{x'} P(v \mid r, x') P(x'). \tag{9}$$

Here $F(y \mid v)$ predicts the ground-truth label y given the representation v. Minimizing this term,

$$\min_{Q,F} V_p(Q, F), \tag{10}$$

corresponds to (1) training the encoder to produce the representation v given the input x and (2) training the predictor to predict the label y given v. The representation v generated by $Q(v \mid x)$ corresponds to the identified causal estimand under the front-door formulation, but we do not claim that this estimand is fully identified in practice.

Second Term $V_d(D, Q)$. The second term is the negative log-likelihood of predicting the domain index k:

$$V_d(D, Q) = \frac{1}{N_{\text{all}}} \sum\nolimits_{(x,y,k)} -\log \sum_v D(k \mid v) Q(v \mid x). \tag{11}$$

The associated minimax game,

$$\min_Q \max_D -\lambda_d V_d(D, Q), \tag{12}$$

trains the discriminator to predict the domain index k given the representation v, and trains the encoder to produce representations that prevent accurate prediction of k.

Training Eqn. 12 to Nash equilibrium enforces marginal independence between k and v, i.e., $k \perp v$, which ensures the required positivity condition discussed in Section 3.2. Combining Eqn. 10 and Eqn. 12 yields Eqn. 3

Which Mechanisms Vary with k and Which Are Invariant. The conditional model $P(y \mid x)$ (and also $P(v \mid x)$) can *vary* with the domain index k. In contrast, under the assumed SCM, the interventional quantity $P(y \mid do(x))$ (and also $P(v \mid do(x))$) and the marginal distribution $P(v)$ are *invariant* with respect to k. Specifically: (1) The invariance of $P(y \mid do(x))$ follows directly from the do-calculus under the assumed SCM (Theorem 3.2), as it represents a causal effect that is independent of the domain index. (2) The invariance of $P(v)$ with respect to k is enforced during training via adversarial learning, corresponding to the min-max optimization in Eqn. 3.

New Insight from Theorem 3.1. Regarding Theorem 3.1, the key insight is not the front-door formula itself, but the **transportability with respect to** K under the assumed SCM, namely,

$$p(v \mid do(x), K = k_1) = p(v \mid do(x), K = k_2), \quad \forall\, k_1 \neq k_2.$$

The expression $p(v \mid do(x))$ happens to be identifiable under the assumed SCM without involving K, which makes it "trivially" transportable across domains; see the definition of *Trivial Transportability* in Definition 3 of (Pearl & Bareinboim, 2011). Theorem 3.1 emphasizes this domain-invariance property rather than re-deriving the front-door criterion.

New Insight from Theorem 3.2. For Theorem 3.2, Eqn. 2 indeed corresponds to the front-door formula under the assumed SCM. However, Eqn. 1 goes one step further by treating $P(v \mid do(x))$ as a computable encoder $Q(v \mid x)$ and using it to compute $p(y \mid do(x))$ via Eqn. 1. Hence, the result is not merely a restatement of the front-door criterion. Moreover, Theorem 3.2 not only establishes identifiability of $p(y \mid do(x))$ under the assumed SCM, but also shows that this expression holds for any domain index k, thereby establishing transportability: $p(y \mid do(x), K = k_1) = p(y \mid do(x), K = k_2), \quad \forall\, k_1 \neq k_2$.

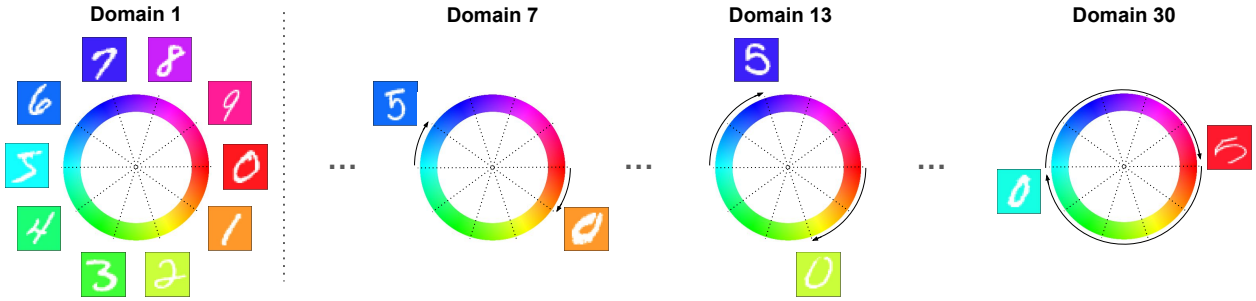


Figure 3: Illustration of the C2MNIST dataset. As shown on the left, the background colors for digit 0 to 9 are evenly separated on the color wheel. As we rotate $6(k-1)^\circ$ clockwise from the base angles (domain 1), the background colors for domain k shift while still maintaining even separation.

Table 2: **C2MNIST accuracy (%) for CADA and various baselines (mean ± std).** We report the accuracy in the source domains and each target domain range. The intervals in the first row represent the domain range of corresponding domains. The average accuracy across target domains is shown in the last column. We use **bold face** to highlight the best results.

Method	$[1,7)$ (Source)	$[7,11)$	$[11,15)$	$[15,19)$	$[19,23)$	$[23,27)$	$[27,31)$	Average
Source-Only	$\mathbf{100.0 \pm 0.0}$	0.1 ± 0.2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
CUA	97.8 ± 1.0	7.2 ± 3.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.2 ± 0.6
ADDA	$\mathbf{100.0 \pm 0.0}$	1.1 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.1
DANN	$\mathbf{100.0 \pm 0.0}$	37.5 ± 20.2	1.7 ± 3.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	6.5 ± 3.7
CDANN	$\underline{99.7 \pm 0.1}$	11.6 ± 9.9	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	0.7 ± 1.5	0.0 ± 0.0	2.0 ± 1.6
MDD	$\mathbf{100.0 \pm 0.0}$	$\underline{59.1 \pm 15.5}$	1.7 ± 1.6	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	10.1 ± 2.9
CIDA	$\mathbf{100.0 \pm 0.0}$	1.4 ± 0.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.1
VOOD	$\mathbf{100.0 \pm 0.0}$	48.0 ± 6.4	$\underline{9.0 \pm 6.1}$	$\underline{8.2 \pm 4.6}$	$\underline{8.0 \pm 4.6}$	$\underline{9.1 \pm 2.7}$	$\underline{9.6 \pm 1.3}$	$\underline{15.3 \pm 2.5}$
GDA	83.3 ± 6.8	3.8 ± 1.5	2.9 ± 0.8	3.4 ± 0.7	3.5 ± 0.8	3.2 ± 0.7	2.4 ± 1.0	3.2 ± 0.8
CADA (Ours)	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{98.7 \pm 2.9}$	$\mathbf{75.6 \pm 17.1}$	$\mathbf{95.7 \pm 3.0}$

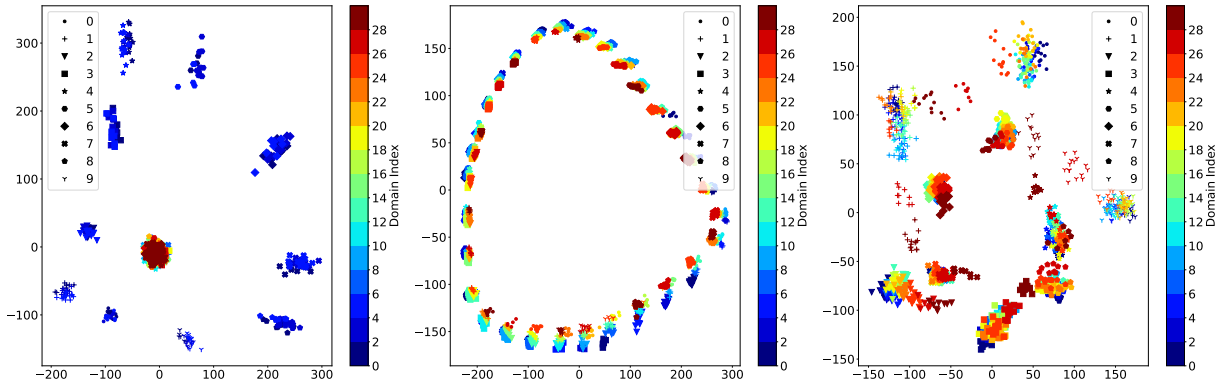
6 Experiments

We evaluated our method, CADA, using a semi-synthetic image dataset, Continuous Colored-MNIST (C2MNIST), along with two real-world medical datasets, Sleep Heart Health Study (SHHS) (Quan et al., 1998) and Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2014), where continuously shifting spurious features are introduced. These empirical studies corroborate the theoretical discoveries outlined earlier and demonstrate the following:

- Using categorical domain adaptation to align continuously indexed domains with continuously shifting spurious features leads to suboptimal alignment and performance.
- Continuously indexed domain adaptation methods alone tend to align shifting spurious features rather than causal features, and are therefore not effective for domain adaptation with continuously shifting spurious features.
- Causality-based domain adaptation methods alone suffer from covariate shift across domains and are not effective in adaptation across continuously indexed domains.
- Our CADA successfully infer and align causal representations from continuously indexed domains even in the presence of continuously shifting spurious features, thereby significantly improving performance.

6.1 Baselines and Implementation Details

We compare CADA with state-of-the-art methods in domain adaptation and causal transportation baselines, including Continuous Unsupervised Adaptation (**CUA**) (Bobu et al., 2018), Adversarial Discriminative Domain Adaptation (**ADDA**) (Tzeng et al., 2017), Domain Adversarial Neural Network (**DANN**) (Ganin et al., 2016), Conditional Domain Adversarial Neural Network (**CDANN**) (Zhao et al., 2017), Margin Disparity Discrepancy (**MDD**) (Zhang et al., 2019), Continuously Indexed Domain Adaptation (**CIDA**) (Wang et al.,



(a) VOOD (b) CIDA (c) CADA

Figure 4: Visualization of learned representations on the C2MNIST dataset, with dimensionality reduction to 2D using principle component analysis (PCA). Domain indices are indicated by color, while data labels are represented by various markers. **(a)** VOOD's source- and target-domain embedding distributions are not aligned due to covariate shift, resulting in poor performance in the target domain. **(b)** CIDA's source- and target-domain embedding distributions are better aligned compared to VOOD. However, they are not aligned by labels. Specifically, these embeddings form 30 clusters, with each cluster containing embeddings with different labels, resulting in poor classification performance. **(c)** CADA's source- and target-domain embedding distributions are aligned by labels. Specifically, these embeddings form 10 clusters, with each cluster containing embeddings with the same label; this makes learning an optimal decision bounary for classification significantly easier, resulting in superior classification performance. See enlarged versions of these figures in Appendix K.

2020), Visual Out-of-Distribution Generalization (**VOOD**) (Mao et al., 2022) (originally proposed for domain generalization and here adapted to the domain adaptation setting with access to unlabeled target data), and Gradual Domain Adaptation (**GDA**) (He et al., 2024).

Since the majority of the baselines are not designed for continuously indexed domains, we made slight generalizations to accommodate these baselines. Specifically, for ADDA, MDD and VOOD, data with different domain indices are merged into one source and one target domain; for DANN, CDANN and CUA, the continuous domain spectrum is discretized into multiple disjoint domains, enabling adaptation between multiple source and target domains. In the case of CUA, the model adapts from the source domains to each target domain individually, progressing from the closest target to the farthest one.

All methods are implemented with PyTorch (Paszke et al., 2019), and run on a single NVIDIA RTX A5000 GPU. For the C2MNIST dataset, all models are trained for 1600 epochs. Our CADA framework's training process is divided into four stages of 400 epochs each, transferring progressively from the previously learned domain to the next: $[7, 13)$, $[13, 19)$, $[19, 25)$ and $[25, 31)$. For healthcare datasets, all models are trained for 50 epochs. Our CADA framework's training process occurs in a single stage, transferring from the source domains to all target domains. We applied a simple grid search to determine the optimal configuration, which is fixed across methods for fair comparison. For C2MNIST and both healthcare datasets, all results in Tables 2, 3, and 4 are reported as mean $\pm$ standard deviation over 5 random seeds.

6.2 Continuous Colored-MNIST

Dataset Description. We start from the simple MNIST (LeCun et al., 1998) dataset since the simplicity of the dataset allows us to study the methods in a controlled environment. We adapt the dataset to continuously indexed domain adaptation with shifting spurious features by adding background colors to the digits, where the color depends on both the domain index and the digit, as illustrated in Fig. 3.

We first create a color wheel from the HSV color space as shown in Fig. 3. For $k = 1$, colors are sampled from $0°, 36°, \ldots, 324°$ of the color wheel for digit $0, 1, \ldots, 9$ respectively. The angles are shown in dotted lines in the figure. For domain $k > 1$, the sampling angles of all the digits are rotated $6(k-1)°$ clockwise from

Table 3: **Accuracy (%) for different methods on the SHHS dataset (mean ± std).** The task is to transfer sleep stage prediction models from patients in the age range $[44, 53)$ to $[53, 90]$. We mark the best result with **bold face** and the second best results with underline.

Method	$[44, 53)$ (Source)	$[53, 58)$	$[58, 63)$	$[63, 68)$	$[68, 73)$	$[73, 78)$	$[78, 83)$	$[83, 90]$	Average
Source-Only	$\mathbf{88.7 \pm 0.2}$	62.9 ± 1.0	61.9 ± 1.2	59.8 ± 1.0	60.3 ± 0.7	60.9 ± 0.9	60.6 ± 0.9	59.0 ± 0.9	60.6 ± 0.6
CUA	85.7 ± 0.4	74.3 ± 1.2	$\underline{75.1 \pm 0.8}$	$\underline{73.2 \pm 0.7}$	$\underline{72.7 \pm 0.8}$	$\underline{70.8 \pm 0.9}$	$\underline{67.6 \pm 1.0}$	$\underline{66.6 \pm 0.9}$	$\underline{71.1 \pm 0.7}$
ADDA	85.3 ± 0.4	$\underline{74.4 \pm 0.2}$	74.7 ± 0.1	72.4 ± 0.2	71.8 ± 0.4	69.9 ± 0.4	66.3 ± 0.7	65.2 ± 0.4	70.2 ± 0.3
DANN	86.2 ± 0.6	72.9 ± 0.7	73.0 ± 0.8	69.8 ± 0.7	67.9 ± 0.7	65.0 ± 0.6	61.3 ± 0.6	60.6 ± 0.8	66.7 ± 0.6
CDANN	85.8 ± 0.6	72.5 ± 0.9	72.7 ± 1.0	69.3 ± 1.1	67.9 ± 1.0	65.3 ± 1.0	61.8 ± 0.8	60.8 ± 0.8	66.7 ± 0.9
MDD	$\underline{88.5 \pm 3.1}$	70.5 ± 0.9	70.1 ± 0.6	70.0 ± 0.5	71.0 ± 0.7	69.2 ± 0.8	66.2 ± 1.2	64.9 ± 1.0	68.5 ± 0.4
CIDA	87.0 ± 1.1	72.0 ± 1.5	71.7 ± 2.3	68.7 ± 2.2	68.4 ± 2.8	68.1 ± 1.2	66.4 ± 0.8	65.4 ± 0.4	68.4 ± 1.2
VOOD	86.9 ± 0.8	60.8 ± 4.8	62.3 ± 3.3	61.5 ± 1.5	61.6 ± 1.4	60.5 ± 2.3	59.4 ± 4.0	58.3 ± 4.3	60.4 ± 1.2
GDA	84.1 ± 2.2	70.8 ± 3.1	71.7 ± 3.1	68.2 ± 3.4	66.1 ± 3.4	62.5 ± 3.7	58.0 ± 3.7	56.2 ± 3.4	64.1 ± 3.3
CADA (Ours)	81.9 ± 1.7	$\mathbf{76.1 \pm 1.2}$	$\mathbf{77.6 \pm 1.1}$	$\mathbf{75.9 \pm 1.0}$	$\mathbf{75.2 \pm 0.4}$	$\mathbf{72.3 \pm 0.4}$	$\mathbf{68.8 \pm 0.8}$	$\mathbf{67.1 \pm 0.5}$	$\mathbf{72.8 \pm 0.3}$

Table 4: **Accuracy (%) for different methods on the MESA dataset (mean ± std).** The task is to transfer sleep stage prediction models from patients in the age range $[54, 59)$ to $[89, 92]$. We mark the best result with **bold face** and the second best results with underline.

Method	$[54, 59)$ (Source)	$[59, 64)$	$[64, 69)$	$[69, 74)$	$[74, 79)$	$[79, 84)$	$[84, 89)$	$[89, 92]$	Average
Source-Only	$\underline{93.6 \pm 0.2}$	36.1 ± 0.4	35.5 ± 0.3	35.8 ± 0.5	41.9 ± 0.6	55.1 ± 1.0	66.1 ± 0.8	56.4 ± 4.3	46.4 ± 0.3
CUA	81.3 ± 3.1	$\underline{74.5 \pm 1.2}$	$\underline{73.8 \pm 0.8}$	$\underline{71.2 \pm 0.4}$	70.8 ± 0.3	70.5 ± 0.6	69.0 ± 1.0	66.0 ± 1.5	$\underline{71.0 \pm 0.3}$
ADDA	86.3 ± 2.7	67.1 ± 2.6	69.9 ± 0.8	69.3 ± 1.8	$\underline{71.0 \pm 1.8}$	$\underline{72.0 \pm 1.4}$	$\underline{69.8 \pm 1.2}$	66.1 ± 2.9	69.4 ± 0.3
DANN	89.0 ± 0.4	58.4 ± 1.8	57.4 ± 1.7	57.3 ± 1.5	60.0 ± 2.1	62.9 ± 2.7	63.4 ± 2.7	64.1 ± 2.3	60.4 ± 1.6
CDANN	86.5 ± 0.4	64.8 ± 2.9	63.7 ± 2.9	61.8 ± 2.7	62.8 ± 2.5	64.3 ± 2.8	63.8 ± 3.0	64.3 ± 2.7	63.6 ± 2.7
MDD	87.4 ± 0.9	60.6 ± 1.5	60.2 ± 1.3	63.1 ± 0.7	67.4 ± 0.2	68.6 ± 1.7	67.9 ± 1.8	$\underline{66.7 \pm 2.1}$	64.9 ± 0.9
CIDA	88.0 ± 1.0	61.4 ± 0.9	60.7 ± 1.0	62.2 ± 1.1	66.0 ± 1.1	67.7 ± 0.8	67.2 ± 0.5	65.7 ± 0.9	64.4 ± 0.3
VOOD	$\mathbf{95.1 \pm 0.2}$	25.4 ± 2.3	26.6 ± 2.9	31.0 ± 3.1	41.6 ± 1.1	59.4 ± 1.6	59.9 ± 2.9	47.3 ± 5.4	41.4 ± 1.6
GDA	73.5 ± 7.8	59.0 ± 4.3	58.8 ± 4.1	57.1 ± 3.9	56.3 ± 3.9	57.8 ± 3.4	54.4 ± 3.2	52.7 ± 5.4	56.7 ± 3.9
CADA (Ours)	80.7 ± 0.6	$\mathbf{74.8 \pm 0.2}$	$\mathbf{74.5 \pm 0.3}$	$\mathbf{73.4 \pm 0.4}$	$\mathbf{74.0 \pm 0.4}$	$\mathbf{74.1 \pm 0.7}$	$\mathbf{73.1 \pm 0.8}$	$\mathbf{71.5 \pm 0.8}$	$\mathbf{73.7 \pm 0.4}$

the base domain $k = 1$. The difference in domain index therefore indicate the distance in the distribution of background colors.

We then treat the domains with $k \leq 6$ as the source domains, and use the remaining ones as the target domains. As a result, for any two distinct digits in the source domains, the ranges their background colors are disjoint, which means that background colors are highly predictive in the source domains. However, these background colors do not generalize to target domains; they are therefore spurious correlations that do not hold across domains. This allows us to verify the empirical performance of different domain adaptation methods.

Accuracy. We further divide the target domains into 6 parts for evaluation based on the distance from the source domains. Table 2 shows the results for different methods. It is evident that in target domains, CUA, ADDA, DANN, and CIDA are completely misled by the continuously shifting spurious features and therefore fail to make predictions in the distant domains. CDANN and MDD, while not entirely wrong in the distant domains, still perform no better than random guessing. While VOOD may theoretically guarantee causal correlation learning, such guarantee break in the presence of covariate shift, which leads to non-positivity in V's probability distributions and non-identifiability of the causal interventional distribution $P(V|do(X))$ (Definition 3.2.3 and Definition 3.2.4 of (Pearl et al., 2000)); see Sec. 3.2 for detailed discussion. As a result, VOOD improves upon random guess only in target domains that are the closest to source domains, i.e., domains $[7, 11)$.

Visualization of Latent-Space Representations. To gain more insights on how CADA (1) aligns representation from different domains to remove covariate shift and (2) eliminates continuously shifting spurious features, we visualize the learned representation for two representative baselines: VOOD and CIDA, as well as our model CADA in Fig. 4. The domain indices are indicated by color, while labels are indicated by various markers.

- **VOOD: Failing to Remove Covariate Shift.** Fig. 4(a) shows VOOD's learned representations, which solely focus on causal correlation learning, unable to align the distributions of source domains (mostly blue points in the figure) and target domains (mostly non-blue points in the figure) due to covariate shift, resulting in poor performance in the target domain (detailed analysis in Sec. 3.2).
- **CIDA: Failing to Remove Continuously Shifting Spurious Features.** Fig. 4(b) shows the representations from CIDA, a typical continuously indexed domain adaptation method. CIDA aligns source- and target-domain embedding distributions better compared to VOOD. However, they are not aligned by labels. Specifically, these embeddings form 30 clusters, with each cluster containing embeddings with different labels, resulting in poor classification performance.
- **CADA: Successfully Removing Both Covariate Shift and Continuously Shifting Spurious Features.** Fig. 4(c) shows the representations from our CADA. CADA successfully align source- and target-domain embedding distributions by labels. Specifically, these embeddings form 10 clusters, with each cluster containing embeddings with the same label; this makes learning an optimal decision bounary for classification significantly easier, resulting in superior classification performance. For better view, please see the enlarged versions of (a), (b), and (c) in Appendix K.

6.3 Healthcare Datasets

Dataset Description. We use two medical datasets, SHHS and MESA to evaluate different methods. Both datasets contain records of subjects' full-night breathing signals and corresponding sleep stage labels for every 30-second segment. Sleep stage labels include "Awake", "Light Sleep N1", "Light Sleep N2", "Deep Sleep", and "Rapid Eye Movement (REM)". Given a breathing signal segment $\mathbf{x}$, a common task in sleep studies is to predict the sleep stage label y. While prior studies may consider 'Light Sleep N1' and 'Light Sleep N2' to be a single stage (Zhao et al., 2017; Wang et al., 2020), we maintain a 5-class classification task in this paper. Among all the subjects' information, age can serve as a natural domain index. The age range of subjects used is $[44, 90]$ and $[54, 92]$ for SHHS and MESA, respectively. Our SHHS and MESA datasets contain continuously shifting spurious noise to simulate the potential increase in noise due to age-related lower-quality sleeping, breathing disorder (e.g., sleep apnea), etc.

Accuracy. Table 3 and Table 4 show the accuracy for different methods on SHHS and MESA, respectively. The target domains of both datasets are divided into 7 parts for evaluation based on the distance from the source domains. One observation is that VOOD, as a causal tranporation method, shows negative improvement compared to Source-Only (e.g., training the model on source domains and directly use it on target domains without adaptation). While methods that directly use categorical domain adaptation perform poorly in the normal continuously indexed domain adaptation setting, CUA shows better performance than other baselines in our setting. On the other hand, our CADA avoids the influence of continuously shifting spurious correlation, demonstrating stable and good performance in both datasets.

6.4 Baselines under Continuous Indices and Incorporating "Age" in Predictive Models

Note that our experiments already include strong continuous-index DA baselines, in particular CIDA and GDA, which are explicitly designed to model continuously indexed domains without relying on discretization or domain merging. These methods therefore constitute appropriate and competitive baselines in the continuous-index setting.

The poor performance of CIDA and GDA in settings with continuously shifting spurious features should be interpreted as revealing the limitations of existing continuous-index DA methods under this form of shift, rather than as a consequence of mismatched problem assumptions or unfair baseline handling.

For discrete domain adaptation baselines (e.g., DANN and CDANN), their failure can be attributed mainly to continuous spurious shift. This is partially supported by the additional empirical results in Appendix H, where changing the bin size in the range of 5$\sim$30 does not significantly affect performance.

Inspired by (Prashant et al., 2025; Reddy et al., 2025), we also explored incorporating the domain index k, i.e., "Age" in the medical datasets, into the predictive models. Our results indicate that explicitly incorporating age

Table 5: Ablation study on the C2MNIST dataset. We compare the full CADA model with its ablated variants: CADA without the causal encoder, CADA without the discriminator, and CADA without causal inference. Accuracy is reported for the source domains and for each target domain range, with intervals in the first row indicating the corresponding domain indices. The final column shows the average accuracy across target domains. The best results are highlighted in **bold**.

Method	$[1, 7)$ (Source)	$[7, 11)$	$[11, 15)$	$[15, 19)$	$[19, 23)$	$[23, 27)$	$[27, 31)$	Average
Source-Only	**100.0**	0.1	0.0	0.0	0.0	0.0	0.0	0.0
CADA w/o Causal Encoder	**100.0**	1.8	0.0	0.0	0.0	0.0	0.0	0.3
CADA w/o Discriminator	**100.0**	48.0	9.9	9.9	9.1	9.6	9.4	16.0
CADA w/o Causal Inference	**100.0**	3.6	0.0	0.0	0.0	0.0	0.0	0.6
CADA (Full)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**82.9**	**97.1**

leads to only marginal changes for CUA and does not substantially improve performance for CADA. CADA consistently outperforms CUA across target age groups, and adding age yields only slight or negligible gains. These findings suggest that CADA mitigates age-induced distributional shifts while preserving predictive structure encoded in physiological signals. For more details, please refer to Appendix I with Tables 10 and 11.

6.5 Ablation Studies

To systematically evaluate the contribution of each major component in CADA, we conduct ablation studies at both the architectural and algorithmic levels.

Architectural Ablations: CADA w/o the Causal Encoder and CADA w/o the Discriminator. In the first variant, i.e., CADA w/o the causal encoder we replace the causal encoder in CADA with a standard encoder to examine the role of causal representation learning. This ablation essentially corresponds to CIDA. In the second variant, i.e., CADA w/o the discriminator, we remove the discriminator from CADA to assess the contribution of adversarial domain alignment.

Algorithmic Ablation: CADA w/o Causal Inference. We also implemented a variant of CADA without causal inference; it uses the causal encoder but optimizes $P(y|x)$ instead of $P(y|do(x))$. This comparison isolates the explicit benefit of causal reasoning in mitigating spurious correlations.

Results. The results are summarized in Table 5. Removing the causal encoder, eliminating the discriminator, or disabling the causal inference mechanism all lead to substantial drops in accuracy. These results indicate that causal representation learning, adversarial domain alignment, and explicit causal reasoning are each essential for robust generalization.

Overall, the ablation experiments demonstrate that both the architectural components and the algorithmic causal inference mechanism provide complementary benefits, and their integration is critical for the robustness and effectiveness of CADA.

7 Conclusion

In this paper, we identify the problem of continuous spurious shift and propose continuously transportable domain adaptation (CADA) as the first general DA method to address this problem. Our theoretical analysis provides insight into how CADA, inspired by causal transportability, encourages transportable representations across continuously indexed domains under certain assumptions. Empirical results on both semi-synthetic and real-world medical datasets show that our method outperforms the state-of-the-art DA methods in the face of continuous spurious shift. Interesting future work includes extending the proposed method to multi-dimensional domain indices, more complex shifting spurious features, and other applications beyond healthcare.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1:107–134, 2013.

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27, 2014.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. 2018.

Peter Bühlmann. Invariance, causality and robustness. 2020.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra Jackson, Michelle Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38, 11 2014. doi: 10.5665/sleep.4732.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020.

Juan D Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, pp. 1661–1667, 2019.

Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Fran**c**cois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1): 2096–2030, 2016.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, Bernhard Schölkopf, et al. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Yifei He, Haoxiang Wang, Bo Li, and Han Zhao. Gradual domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 25(361):1–40, 2024. URL `http://jmlr.org/papers/v25/23-1180.html`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Tianyi Liu, Zihao Xu, Hao He, Guang-Yuan Hao, Guang-He Lee, and Hao Wang. Taxonomy-structured domain adaptation. In *International Conference on on Machine Learning*, 2023.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3947–3956, 2021.

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *CVPR*, 2022.

Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, and Eric Granger. Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1347, 2021.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *TNN*, 22(2):199–210, 2010.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 247–254, 2011.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8558–8567, 2021.

Parjanya Prashant, Seyedeh Baharan Khatami, Bruno Ribeiro, and Babak Salimi. Scalable out-of-distribution robustness in the presence of unobserved confounders, 2025.

Stuart Quan, Barbara Howard, Conrad Iber, James Kiley, F. Nieto, George O'Connor, David Rapoport, Susan Redline, John Robbins, Jonathan Samet, and ‡Patricia Wahl. The sleep heart health study: Design, rationale, and methods. *Sleep*, 20:1077–85, 01 1998. doi: 10.1093/sleep/20.12.1077.

Abbavaram Gowtham Reddy, Celia Rubio-Madrigal, Rebekka Burkholz, and Krikamol Muandet. When shift happens - confounding is to blame, 2025.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *ICCV workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, pp. 443–450, 2016.

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

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pp. 7167–7176, 2017.

Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *ICML*, 2020.

Zihao Xu, Hao He, Guang-He Lee, Yuyang Wang, and Hao Wang. Graph-relational domain adaptation. In *ICLR*, 2022.

Zihao Xu, Guang-Yuan Hao, Hao He, and Hao Wang. Domain-indexing variational bayes: Interpretable domain index for domain adaptation. In *International Conference on Learning Representations*, 2023.

Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8599–8608, 2021.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I Jordan. Bridging theory and algorithm for domain adaptation. *arXiv preprint arXiv:1904.05801*, 2019.

Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S. Jaakkola, and Matt T. Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *ICML*, pp. 4100–4109, 2017.

Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

A Discussion on Structural Causal Modeling Assumptions

The structural causal model introduced in this paper serves as a high-level abstraction to reason about identifiability, transportability, and invariance under a causality-inspired formulation. It is not intended as a literal description of neuron-level mechanisms, but rather as a representation of latent semantic factors that the learned encodings aim to approximate.

Under this abstraction, the validity of the formulation depends on whether the learned representations meaningfully reflect such latent factors and approximately respect the assumed structural relationships. As in other causality-inspired representation learning approaches, this assumption may be challenged when representations strongly entangle multiple factors or when the data-generating process deviates from the assumed structural relationships.

The results in this paper therefore establish transportability guarantees under the stated structural assumptions. When these assumptions are approximately satisfied, the proposed formulation provides a principled way to remove continuously shifting spurious correlations. When they are violated, the method may lose its theoretical guarantees, as is typical in causality-motivated modeling frameworks.

B Inference Algorithm of CADA

Algorithm 1 shows details for CADA's inference (prediction) process.

Algorithm 1 Inference (Prediction) Using CADA

Input: Query x, data distribution $\mathcal{D}$ over $\{(x, y, k)\}$, causality-inspired encoder $Q(v|x)$ consisting of the intermediate encoder $P(r|x)$ and the augmented encoder $P(v|r, x')$. Numbers of samples for r, x', and v, i.e., N_r, N_x, and N_v.
for $i = 1, ..., N_r$ **do**
  Sample $r^{(i)} \sim P(r|x)$.
  for $j = 1, ..., N_x$ **do**
    Sample from the data distribution: $\mathbf{x}'^{(ij)} \sim P(x')$.
    for $m = 1, ..., N_v$ **do**
      Sample $v^{(ijm)} \sim P(v|r^{(i)}, \mathbf{x}'^{(ij)})$.
    end for
  end for
end for
Compute the interventional quantity effect for each class: $P(y|do(x)) = \frac{1}{N_r N_x N_v} \sum_{i=1}^{N_r} \sum_{j=1}^{N_x} \sum_{m=1}^{N_v} P(y|v^{(ijm)})$.
Output: Class prediction $\widehat{y} = \operatorname{argmax}_y P(y|do(x))$.

C Definitions

Below we provide the formal definitions of spurious association and spurious shift in the context of CADA.

Definition 2 (Spurious Association)**.** *Two variables X and Y are spuriously associated if they are dependent in some context and there exist two other variables ($Z_1$and Z_2) and two contexts (S_1 and S_2) such that:*

1. *Z_1 and X are dependent given S_1 (i.e. $Z_1 \not\perp\!\!\!\perp X|S_1$)*
2. *Z_1 and Y are independent given S_1 (i.e. $Z_1 \perp\!\!\!\perp Y|S_1$)*
3. *Z_2 and Y are dependent given S_2 (i.e. $Z_2 \not\perp\!\!\!\perp Y|S_2$)*
4. *Z_2 and X are independent given S_2 (i.e. $Z_2 \perp\!\!\!\perp X|S_2$)*

Definition 3 (Spurious Shift)**.** *Given a latent representation $r = (r_c, r_s)$, where r_c captures the causal factors for predicting y, and r_s captures spurious features, we say a spurious shift occurs when:*

$$P_{source}(y|r_c) = P_{target}(y|r_c), P_{source}(y|r_s) \neq P_{target}(y|r_s)$$

That is, the causal mechanism is invariant, but the spurious correlations shift across domains.

D Computational Cost

Table 6 shows the time cost for different models.

Table 6: Accuracy (%) and Time Cost (Milliseconds per Image) for Different Numbers of Samples.

Model	# Samples	[1, 7) (Source)	[7, 11)	[11, 15)	[15, 19)	[19, 23)	[23, 27)	[27, 31)	Average	Time
VOOD (Baseline)	-	100.0	48.0	9.9	9.9	9.1	9.6	9.4	16.0	0.40 ms
CIDA (Baseline)	-	100.0	1.8	0.0	0.0	0.0	0.0	0.0	0.3	0.13 ms
CADA	1	100.0	100.0	100.0	100.0	100.0	99.9	74.7	95.8	0.29 ms
CADA	4	100.0	100.0	100.0	100.0	100.0	100.0	76.5	96.1	0.31 ms
CADA	9	100.0	100.0	100.0	100.0	100.0	100.0	77.2	96.2	0.35 ms
CADA (Original)	55	100.0	100.0	100.0	100.0	100.0	100.0	82.9	97.1	0.41 ms

Table 7: **Average accuracy, Macro-F1, and per-class F1 scores on the SHHS dataset.** The best results are highlighted in **bold**.

Method	Avg Acc	Macro-F1	Awake	Light Sleep	Deep Sleep	REM
Source-Only	60.6	62.9	68.3	62.0	49.1	72.2
CUA	71.1	71.1	83.5	73.4	53.9	73.8
ADDA	70.2	70.8	81.3	72.3	55.0	74.6
DANN	66.7	68.8	77.4	67.5	55.5	74.8
CDANN	66.7	68.5	74.8	67.5	**56.1**	75.6
MDD	68.5	68.6	79.4	70.7	48.5	75.8
CIDA	68.4	70.0	80.3	70.4	55.1	74.3
VOOD	60.4	64.0	64.8	66.4	52.0	72.7
GDA	64.1	62.1	67.9	**74.2**	49.4	57.1
CADA (Ours)	**72.8**	**72.0**	**84.5**	70.7	**56.1**	**76.7**

E Additional Metrics on Medical Datasets

To provide a more comprehensive evaluation, we report average accuracy, macro-F1, and per-class F1 scores on the SHHS dataset. Results are summarized in Table 7.

F More Implementation Details

Compute Resources. All methods are implemented with PyTorch (Paszke et al., 2019), and run on a single NVIDIA RTX A5000 GPU.

Training Process. For the C2MNIST dataset, all models are trained for 1600 epochs. Our CADA framework's training process is divided into four stages of 400 epochs each, transferring progressively from the previously learned domain to the next: $[7, 13)$, $[13, 19)$, $[19, 25)$ and $[25, 31)$. For healthcare datasets, all models are trained for 50 epochs. Our CADA framework's training process occurs in a single stage, transferring from the source domains to all target domains. We applied a simple grid search to determine the optimal configuration, which is fixed across methods for fair comparison. For C2MNIST, we tune the adversarial weight λ_d over $\{0.5, 1, 3, 5, 7, 8, 10\}$. For sleep datasets, λ_d is selected from $\{0.25, 0.5, 0.75, 1, 2, 4, 6, 8, 10, 12\}$. For methods involving dropout (e.g., VOOD and CADA), the dropout rate is selected from $\{0.55, 0.65, 0.75, 0.85, 0.95\}$.

Model Architectures. For fair comparison, we adopt the same backbone neural network architectures for baseline methods and CADA within the same dataset. In C2MNIST, we use the same multi-layer perceptron as in CIDA (Wang et al., 2020), with the hidden dimension set to 800. For healthcare datasets, we adopt the same setting as in CIDA for sleep learning, with the hidden dimension set to 384.

Semi-Synthetic Construction for Sleep Data. For the sleep datasets (SHHS and MESA), continuously shifting spurious correlations are introduced in a semi-synthetic manner. Specifically, we first extract intermediate representations from a pretrained sleep-stage encoder, and then inject additive Gaussian noise whose magnitude varies smoothly with the continuous domain index (age) and interacts with the class label. This construction induces a continuous change in $P(X \mid Y)$ across age groups while keeping the underlying physiological structure unchanged.

G Sampling Strategies for x' and Sensitivity Analysis

In our implementation, we adopt stage-wise sampling aligned with the progressive structure of continuous adaptation. Specifically, continuous adaptation is divided into multiple stages along the domain index, and

at each stage x' is sampled from the data of the immediately preceding stage (including source data and unlabeled target data).

To study the sensitivity of the method to different x'-sampling schemes, we further evaluate the following strategies:

(1) **Source-only sampling**, where x' is drawn exclusively from the source domain.

(2) **Stage-wise sampling**, where x' is sampled from the data of the immediately preceding stage.

(3) **Cumulative sampling**, where x' is sampled from all source data together with unlabeled target data from previous stages.

All samples of x' are drawn according to the natural data distribution within the selected domains, without conditioning on labels. Pseudo-labels are not used, and no class-balanced sampling is performed.

Table 8 summarizes the results on C2MNIST.

Table 8: Accuracy across different x'-sampling strategies on C2MNIST.

Sampling Method	[1,7) (Source)	[7,11)	[11,15)	[15,19)	[19,23)	[23,27)	[27,31)	Avg
Source-only	100.0	100.0	100.0	100.0	100.0	100.0	72.5	96.1
Stage-wise	100.0	100.0	100.0	100.0	100.0	100.0	72.5	96.1
Cumulative	100.0	100.0	100.0	100.0	100.0	100.0	72.5	96.1

We observe that all three sampling strategies yield nearly identical performance across domain groups, indicating that the method is largely insensitive to the specific choice of x'-sampling scheme.

If the samples of x' are extremely label-imbalanced, expectations over x' may be dominated by majority classes, which can degrade performance on minority classes. However, the causal functional itself does not require label-balanced sampling; the integration over x' is intended to marginalize domain-specific variations rather than to enforce class balance.

H Sensitivity to Domain Index Discretization

We conducted additional experiments to examine the sensitivity of discretized continuous-index baselines to different binning strategies.

The original C2MNIST setup contains 30 ordered domains (6 source and 24 target domains). The 30-bin setting corresponds to the original per-domain discretization, where each domain is treated as a distinct bin.

To evaluate the effect of coarser discretization, we further consider 5, 10, and 15 bins, obtained by merging neighboring domains into progressively larger bins. Together with the 30-bin setting, these choices span a reasonable range of granularity.

For DANN and CDANN, we report results under all four bin sizes (5, 10, 15, and 30 bins). The results are summarized in Table 9. These results show that changing the bin size (or the number of bins) does not significantly affect the poor performance of these discrete domain adaptation baselines.

I Additional Experiments Incorporating Age

Recent work has argued that, in certain settings, incorporating observed confounders into predictive models may be preferable to enforcing invariance (Prashant et al., 2025; Reddy et al., 2025).

In our healthcare experiments, age is treated as the domain index K. While age may contain predictive information, any age-related signal that manifests in physiological measurements is already reflected in the

Table 9: Accuracy (%) with different bin sizes for discretized baselines on the C2MNIST dataset.

Method	Bin Size	$[1, 7)$	$[7, 11)$	$[11, 15)$	$[15, 19)$	$[19, 23)$	$[23, 27)$	$[27, 31)$	Average
DANN	5 bins	100.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3
DANN	10 bins	100.0	23.4	0.0	0.0	0.0	0.0	0.0	17.6
DANN	15 bins	100.0	23.6	0.0	0.0	0.0	0.0	0.0	17.7
DANN	30 bins	100.0	58.3	8.9	0.0	0.0	0.0	0.0	23.9
CDANN	5 bins	100.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3
CDANN	10 bins	99.8	11.1	0.0	0.0	0.0	0.0	0.0	15.8
CDANN	15 bins	100.0	35.0	4.3	0.0	0.0	0.0	0.0	19.9
CDANN	30 bins	99.6	5.5	0.1	0.1	0.0	0.0	0.0	15.0

Table 10: Additional experiments incorporating age on SHHS.

Method	[44,53)	[53,58)	[58,63)	[63,68)	[68,73)	[73,78)	[78,83)	[83,90]	Avg
CUA	85.7	74.3	75.1	73.2	72.7	70.8	67.6	66.6	71.1
CUA + Age	85.8	74.4	75.0	72.9	72.6	71.2	68.0	67.5	71.3
CADA	81.9	76.1	77.6	75.9	75.2	72.3	68.8	67.1	72.8
CADA + Age	85.1	76.6	77.5	75.9	74.9	72.4	69.6	66.7	72.9

observed breathing signals X. For example, older patients may exhibit more frequent "Awake" periods. Such patterns correspond to irregular breathing signals and can be learned directly from X without explicitly incorporating age as an input feature.

To examine whether explicitly incorporating age improves predictive performance, we conduct additional experiments where age is concatenated to the input representation. We compare CADA against the strongest baseline in our setting (CUA), and also evaluate the corresponding "+ Age" variants. Results on SHHS and MESA are shown in Tables 10 and 11, respectively.

The results indicate that explicitly incorporating age leads to only marginal changes for CUA and does not substantially improve performance for CADA. CADA consistently outperforms CUA across target age groups, and adding age yields only slight or negligible gains. These findings suggest that CADA mitigates age-induced distributional shifts while preserving predictive structure encoded in physiological signals.

J Scope, Limitations, and Extensions

The proposed framework operates on learned representations and does not rely on modality-specific assumptions. In this work, we evaluate the method on two substantially different modalities, i.e., images (C2MNIST) and medical time-series data (SHHS and MESA), demonstrating that the same causality-inspired formulation applies across modalities.

More broadly, the method can be extended to other modalities, such as text or videos, by instantiating the encoder with modality-appropriate architectures while keeping the causality-inspired modeling and domain-invariance objectives unchanged.

K Enlarged Figures for Representation Visualization

Fig. 5, Fig. 6, and Fig. 7 are the enlarged versions of Fig. 4(a), Fig. 4(b), and Fig. 4(c) in the main paper.

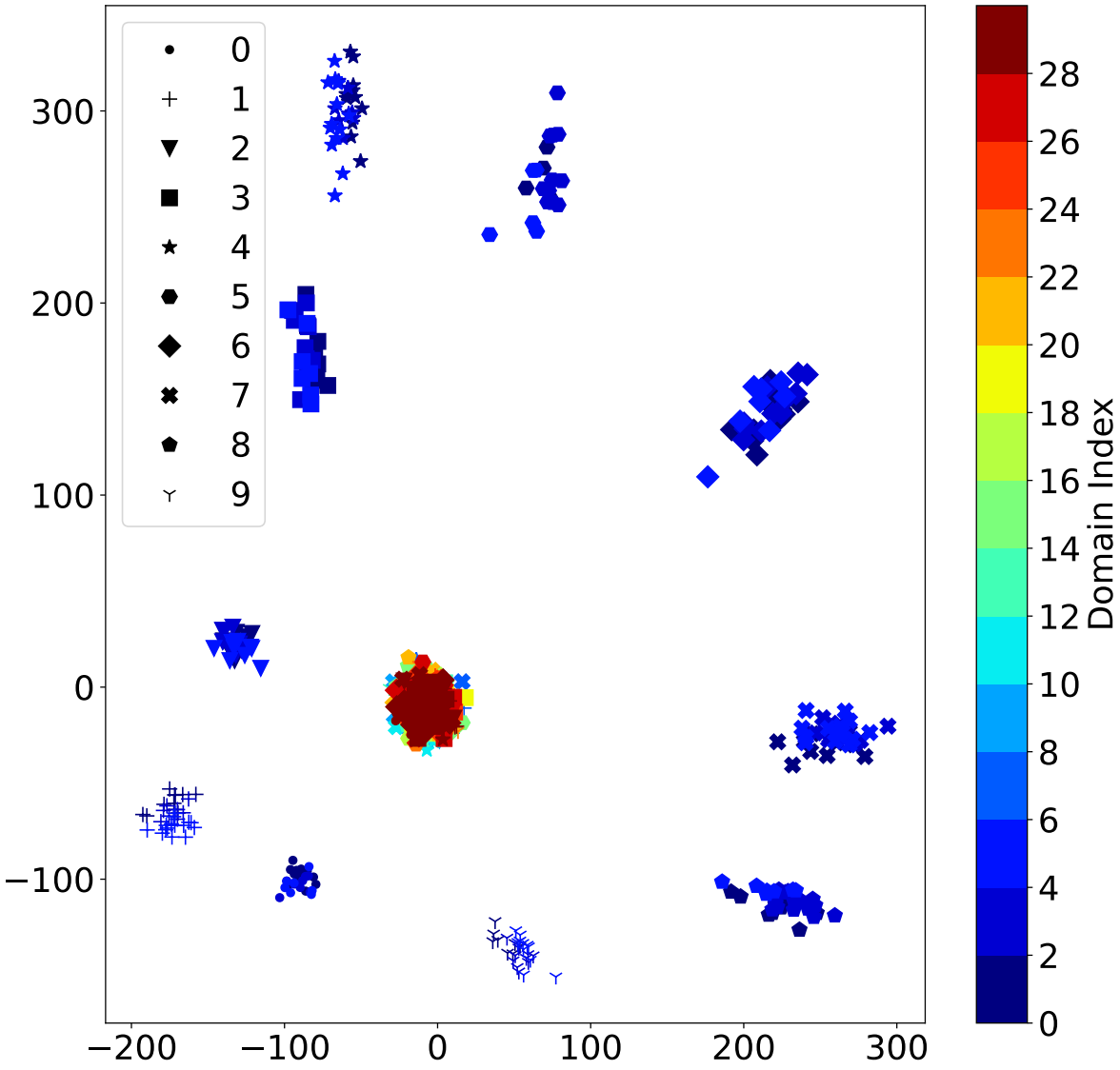


Figure 5: Enlarged version of Fig. 4(a): Visualization of VOOD's representations.

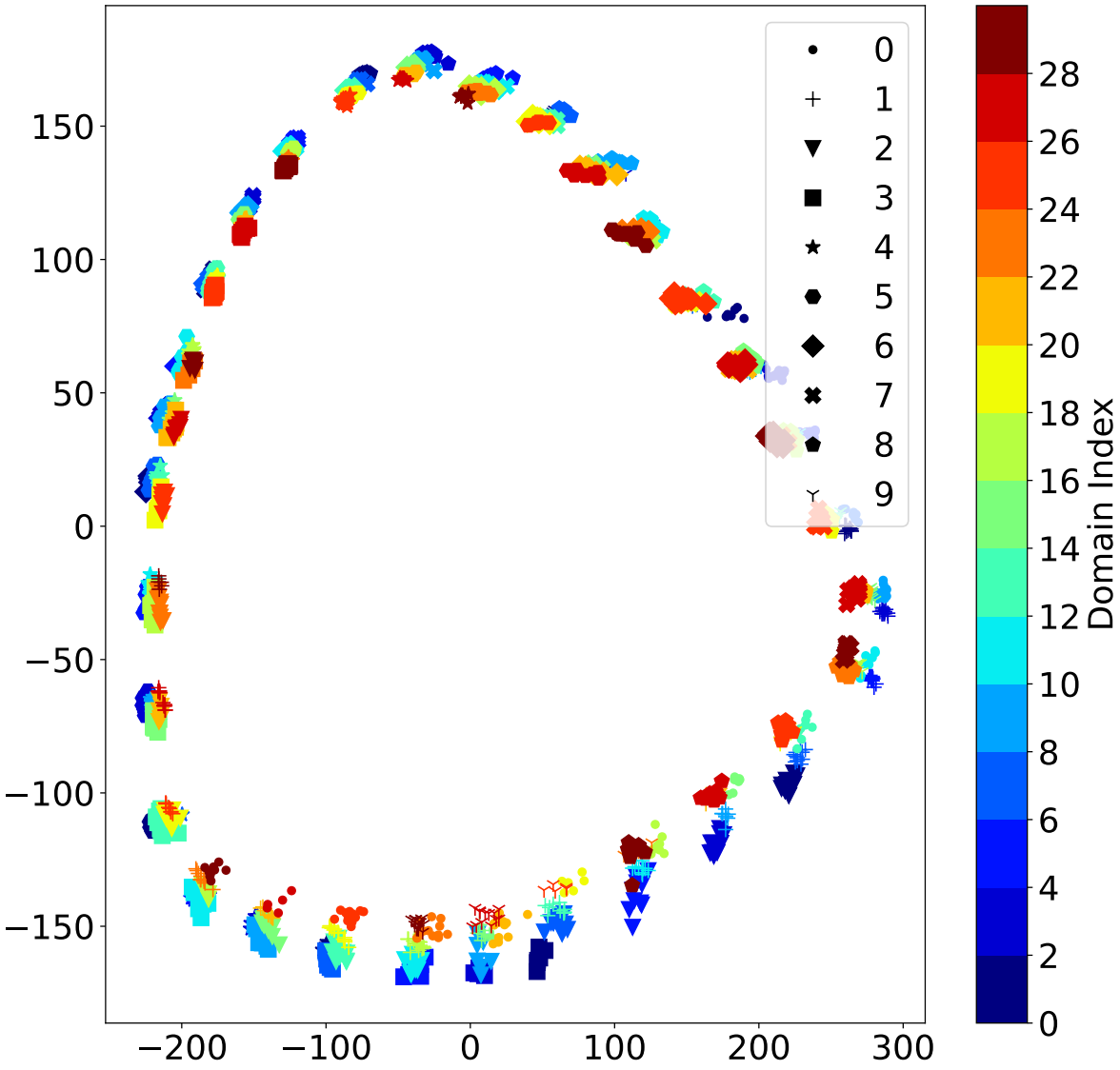


Figure 6: Enlarged version of Fig. 4(b): Visualization of CIDA's representations.

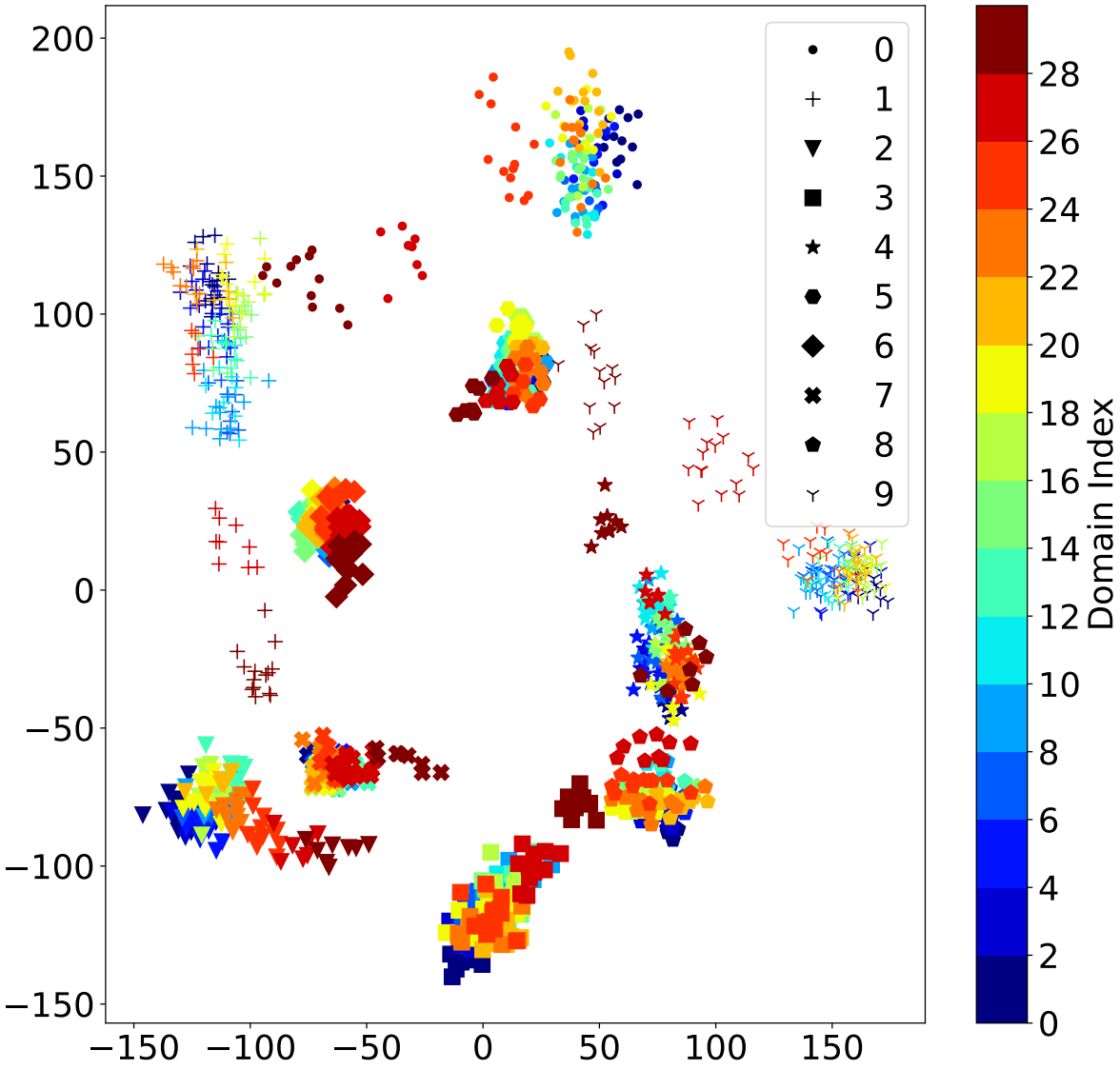


Figure 7: Enlarged version of Fig. 4(c): Visualization of CADA's representations.

Table 11: Additional experiments incorporating age on MESA.

Method	[54,59)	[59,64)	[64,69)	[69,74)	[74,79)	[79,84)	[84,89)	[89,92]	Avg
CUA	81.3	74.5	73.8	71.2	70.8	70.5	69.0	66.0	71.0
CUA + Age	80.1	75.1	74.1	71.2	70.6	70.2	68.1	65.1	70.8
CADA	80.7	74.8	74.5	73.4	74.0	74.1	73.1	71.5	73.7
CADA + Age	80.8	74.5	74.2	73.0	73.8	74.2	72.3	67.3	72.9