CONDITION COVERAGE IN SPARSE TARGET CALIBRATION FOR DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study domain adaptation when target labels are available only at a sparse set of calibration levels, while prediction is required across the full target label space. We show that the observed target distributions do not uniquely determine the target distributions at unseen levels, even when the observed distributions are known exactly. This motivates a distributionally robust formulation that accounts for uncertainty induced by incomplete calibration coverage. Based on the resulting robust bound, we propose ***CaDRA***, a **Ca**libration-aware **D**istributionally **R**obust **A**daptation method that combines labeled target prediction with balanced source–target alignment. Controlled synthetic experiments and three real-world transfer tasks show consistent performance across different calibration patterns and domain shifts.

1 INTRODUCTION

Domain adaptation studies how to transfer a model trained on a source domain to a target domain with a different data distribution (Ben-David et al., 2010). In practice, source and target data are often collected from different machines, sensors, or environments, so sensor observations corresponding to the same underlying physical condition may differ across domains. Existing methods commonly assume access to target samples that broadly cover the target distribution, either unlabeled or through a small labeled target set (Long et al., 2015; Ganin et al., 2016; Sun & Saenko, 2016; Teshima et al., 2020).

In many industrial systems, the target label represents a physical condition to be predicted, such as fault severity (Lessmeier et al., 2016; Zhang et al., 2023). Healthy operation can usually be measured repeatedly, whereas collecting labeled data at additional severity levels often requires costly controlled experiments (Zhang et al., 2023). As a result, target data may be abundant at a few levels but unavailable at others. We study this setting as *sparse target calibration for domain adaptation*. Labeled target data are available only at a subset of label levels, while prediction is required across the full target label space.

This setting differs from few-shot domain adaptation, which considers only a small number of labeled target samples (Teshima et al., 2020). Here, target data may be abundant at the observed levels, while other label levels are entirely unobserved during adaptation. The corresponding target conditional distributions therefore cannot be estimated directly from target data. Meanwhile, source and target conditional distributions may still differ at the observed levels (Long et al., 2015; Ganin et al., 2016; Sun & Saenko, 2016; Zhang & Fink, 2025). To generalize across the full target label space, an adaptation method must handle source–target discrepancy at observed levels and distributional uncertainty at unseen levels.

We address this problem through distributionally robust optimization (DRO) (Mohajerin Esfahani & Kuhn, 2018; Sinha et al., 2018). We first show that target data observed at a subset of label levels do not uniquely determine the target distributions at unseen levels, even when the observed distributions are known exactly. We then quantify the uncertainty at unseen levels using the available target information and formulate a robust adaptation problem over the full target label space. Based on this analysis, we propose *CaDRA*, a **Ca**libration-aware **D**istributionally **R**obust **A**daptation method that combines labeled target prediction with source–target alignment.

Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- We show that the partially observed target distributions do not uniquely determine the target distributions at unseen levels. This non-identifiability motivates a distributionally robust formulation for domain adaptation with sparse target calibration.
- We derive a tractable empirical surrogate for the resulting DRO problem, specialized to domain adaptation. CaDRA combines labeled target prediction with source–target alignment and can be optimized using only the available source and target data.
- We evaluate CaDRA on controlled synthetic shifts and three real-world transfer settings, including cross-machine, cross-sensor, and cross-process adaptation. CaDRA performs consistently across different coverage and transfer regimes.

2 Related Work

Domain adaptation. Classical domain adaptation methods reduce source–target mismatch through adversarial invariance, moment matching, or discrepancy minimization. DANN learns features that confuse a domain discriminator (Ganin et al., 2016); CORAL aligns second-order feature statistics (Sun & Saenko, 2016); and DAN uses kernel mean embeddings to reduce distribution discrepancy (Long et al., 2015). These approaches provide important generic baselines. However, they were not designed for target supervision restricted to a sparse set of regression conditions. Regression requires preserving metric structure in the output and representation spaces. RSD studies representation-subspace geometry for deep domain adaptation regression (Chen et al., 2021). COD develops a conditional discrepancy specifically for continuous-output domain adaptation (Yang et al., 2024). More recently, LGOT incorporates label geometry into optimal-transport alignment (Chen et al., 2026). Our setting differs in supervision structure, i.e., target samples are available only at controlled calibration conditions, while entire target conditions remain unavailable during adaptation.

Sparse supervised adaptation. Semi-supervised and few-shot domain adaptation study transfer when limited labeled target data are available (Saito et al., 2019; Jiang et al., 2020). For regression, Teshima et al. (2020) study few-shot adaptation through causal mechanism transfer. These settings primarily limit the number of labeled target samples. Our setting differs in what is sparse: many target samples may be available at each observed level, while entire levels of the target label space remain unobserved during adaptation. Increasing the number of samples at an observed level can improve estimation there, but does not provide target information at missing levels.

Distributionally robust optimization. Distributionally robust optimization minimizes worst-case risk over an ambiguity set of plausible data distributions. Wasserstein ambiguity sets have been widely studied because they provide a flexible notion of distributional perturbation together with tractable reformulations (Gao & Kleywegt, 2023; Mohajerin Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019). They have also been used in statistical learning and robust prediction (Sinha et al., 2018; Chen & Paschalidis, 2018; Kuhn et al., 2019). In our setting, the uncertainty arises from a different source: target conditional distributions are observed only at a subset of label levels, while the distributions at the remaining levels are unavailable. We use Wasserstein ambiguity sets to represent this missing target information and derive the calibration-and-alignment objective used in CaDRA.

3 Problem Formulation

We study domain adaptation for predicting a physical quantity from sensor measurements collected in different domains. Let $C \in \mathcal{C}$ denote the quantity to be predicted, $X \in \mathcal{X}$ the corresponding sensor measurement, and $M \in \mathcal{M}$ the observed operating variables available during both adaptation and deployment, where $\mathcal{C} = \{c_1, \ldots, c_K\} \subset \mathbb{R}$. We use uppercase letters for random variables and lowercase letters for their realizations. For example, in motor fault analysis, C may represent fault severity, X may contain vibration measurements, and M may include speed and load.

The source domain provides labeled data across all levels in $\mathcal{C}$ and a pretrained predictor $f_s(x) = H_s(E_s(x))$, where E_s is the source encoder and H_s is the prediction head. The target domain corresponds to a new deployment domain, where labeled data are available only at a subset of levels

in $\mathcal{C}$. Our goal is to adapt the pretrained source model using these labeled target data so that it can predict C across the full target label space.

Let P_s and P_t denote the joint source and target distributions over (X, M, C). The two domains share the same physical levels in $\mathcal{C}$, but differ in the measurement domain, such as the machine, sensor configuration, or data-generation process. Consequently, even at the same physical level and operating state, the sensor measurements may follow different conditional distributions:

$$P_s(X \mid C = c, M = m) \neq P_t(X \mid C = c, M = m).$$

During target adaptation, the source encoder $E_s : \mathcal{X} \to \mathcal{Z}$ and prediction head $H_s : \mathcal{Z} \to \mathbb{R}$ remain fixed, while a target adapter $T_\phi : \mathcal{Z} \times \mathcal{M} \to \mathcal{Z}$ is learned. The adapted target predictor is $f_t^\phi(x, m) = H_s(T_\phi(E_s(x), m))$. Thus, f_s is the pretrained source predictor, whereas f_t^ϕ applies the learned target adapter before the fixed prediction head. We denote the source latent representation by $Z_s = E_s(X)$ for $(X, M, C) \sim P_s$, and the adapted target representation by $Z_t^\phi = T_\phi(E_s(X), M)$ for $(X, M, C) \sim P_t$.

Let $\mathcal{A} \subset \mathcal{C}$ denote the set of physical levels at which labeled target data are available, and let $\mathcal{U} = \mathcal{C} \setminus \mathcal{A}$ denote the remaining unseen levels. No target samples from $\mathcal{U}$ are used during adaptation. Sparse target calibration therefore refers to incomplete coverage of the target label space, rather than to a small number of observations at each observed level. This setting also includes the extreme case where only one physical level, such as the healthy condition, is observed in the target domain.

For each physical level $c \in \mathcal{C}$, define the conditional target MAE as $\mathcal{R}_t(c; \phi) = \mathbb{E}_{(X,M)\sim P_t(\cdot,\cdot|C=c)}[|f_t^\phi(X, M) - c|]$. The full-target risk is

$$\mathcal{R}_t^{\text{all}}(\phi) = \sum_{c\in\mathcal{C}} \pi_c \mathcal{R}_t(c; \phi), \qquad \pi_c \geq 0, \quad \sum_{c\in\mathcal{C}} \pi_c = 1. \tag{1}$$

When deployment frequencies are unavailable, we set $\pi_c = 1/|\mathcal{C}|$, so that all physical levels receive equal weight.

Ideally, target adaptation would minimize $\mathcal{R}_t^{\text{all}}(\phi)$ over the adapter parameter ϕ. However, target data are unavailable at levels in $\mathcal{U}$, so the corresponding conditional risks cannot be evaluated during adaptation.

4 Distributionally Robust Optimization (DRO) Formulation

Target labels are available only at calibration levels $\mathcal{A}$, while prediction is required over the full set $\mathcal{C}$. The remaining levels $\mathcal{U} = \mathcal{C} \setminus \mathcal{A}$ are unobserved during adaptation, and their target conditional distributions cannot in general be uniquely determined from the observed calibration levels.

We first introduce the conditional latent distributions and the regularity assumptions used in the analysis. We then show that the observed calibration levels do not uniquely determine the target distributions at unseen levels and use this uncertainty to construct conditional ambiguity sets. Optimizing the worst-case risk over these sets yields a robust target-risk bound in terms of calibration prediction error, source–target discrepancy, and calibration coverage.

4.1 Conditional Distributions and Cross-Level Regularity

For each physical level $c \in \mathcal{C}$, let $P_s^c = \mathcal{L}(Z_s \mid C = c)$ and $Q_\phi^c = \mathcal{L}(Z_t^\phi \mid C = c)$ denote the corresponding source and adapted target conditional latent distributions. We measure their source–target discrepancy by

$$\Delta_\phi(c) = W_1(Q_\phi^c, P_s^c), \tag{2}$$

where W_1 denotes the 1-Wasserstein distance on the latent space.

We assume that the source prediction head H_s is L_H-Lipschitz on the latent space, i.e., $|H_s(z) - H_s(z')| \leq L_H\|z - z'\|_2$ for all $z, z' \in \mathcal{Z}$. We further assume that the source and adapted target conditional distributions vary smoothly across physical levels. For some $L_s, L_t \geq 0$, and all adapters ϕ considered,

$$W_1(P_s^c, P_s^{c'}) \leq L_s|c - c'|, \qquad W_1(Q_\phi^c, Q_\phi^{c'}) \leq L_t|c - c'|. \tag{3}$$

Here, L_H controls the sensitivity of prediction to latent perturbations, whereas L_s and L_t control cross-level variation of the conditional latent distributions.

4.2 Why Unseen Target Distributions Are Ambiguous

We first show that the target distributions at unseen levels cannot, in general, be uniquely recovered from calibration information under the regularity assumptions above.

Let $\mathfrak{Q}(L_t)$ denote the class of target conditional distribution families $\mathbf{Q} = (Q^c)_{c\in\mathcal{C}}$ satisfying the target-side regularity condition in Eq. (3). We use $a \in \mathcal{A}$ to denote a calibration level and $u \in \mathcal{U}$ to denote an unseen level. To isolate the effect of calibration coverage, we consider an oracle setting in which the target latent conditional distributions Q^a are known exactly for all $a \in \mathcal{A}$. We also allow the learner to use the complete source distribution information, while no target information is available at unseen levels. Let $\widehat{\mathbf{Q}} = (\widehat{Q}^u)_{u\in\mathcal{U}}$ denote any estimate of the unseen target distributions based on this information.

For each $u \in \mathcal{U}$, let $a(u) \in \arg\min_{a\in\mathcal{A}} |u - a|$ denote a nearest calibration level and define

$$\rho(\mathcal{A}) = \sum_{u\in\mathcal{U}} \pi_u |u - a(u)|.$$

Theorem 1 (Calibration-coverage lower bound)**.**

$$\inf_{\widehat{\mathbf{Q}}} \sup_{\mathbf{Q}\in\mathfrak{Q}(L_t)} \sum_{u\in\mathcal{U}} \pi_u W_1(\widehat{Q}^u, Q^u) \geq L_t \rho(\mathcal{A}). \tag{4}$$

The proof is given in Appendix D. Even with exact knowledge of all target distributions at calibration levels, the target distributions at unseen levels cannot be recovered arbitrarily accurately. Their worst-case estimation error is at least $L_t\rho(\mathcal{A})$. Thus, sparse calibration leaves an unavoidable set of plausible target distributions at unseen levels rather than a uniquely identifiable target distribution.

4.3 Conditional Ambiguity Sets

The preceding result motivates a set-valued representation of the unknown target distribution at each unseen level. Under Eq. (3) and the triangle inequality,

$$\Delta_\phi(u) \leq \Delta_\phi(a(u)) + (L_s + L_t)|u - a(u)|.$$

The proof is given in Appendix A. For a fixed adapter ϕ, define

$$r_\phi(u) = \Delta_\phi(a(u)) + (L_s + L_t)|u - a(u)|$$

and the conditional ambiguity set

$$\mathbb{B}_\phi(u) = \{Q : W_1(Q, P_s^u) \leq r_\phi(u)\}. \tag{5}$$

By construction, $Q_\phi^u \in \mathbb{B}_\phi(u)$.

At calibration levels, the target risk can be evaluated directly. At unseen levels, the true target distribution is unknown but belongs to $\mathbb{B}_\phi(u)$. We therefore upper-bound the target risk at u by the worst-case risk over $\mathbb{B}_\phi(u)$. The resulting robust full-target risk is

$$\mathcal{R}_t^{\text{robust}}(\phi) = \sum_{a\in\mathcal{A}} \pi_a \mathcal{R}_t(a;\phi) + \sum_{u\in\mathcal{U}} \pi_u \sup_{Q\in\mathbb{B}_\phi(u)} \mathbb{E}_{Z\sim Q}|H_s(Z) - u|. \tag{6}$$

4.4 Robust Full-Target Upper Bound

Theorem 2 (Robust full-target-risk bound)**.** *Under the regularity conditions above,*

$$\mathcal{R}_t^{\text{all}}(\phi) \leq \mathcal{R}_t^{\text{robust}}(\phi) \leq \underbrace{\sum_{a\in\mathcal{A}} \alpha_a \mathcal{R}_t(a;\phi)}_{\textit{calibration prediction}} + 2L_H \underbrace{\sum_{a\in\mathcal{A}} \beta_a \Delta_\phi(a)}_{\textit{calibration discrepancy}} + \underbrace{\kappa_{\text{cov}}\rho(\mathcal{A})}_{\textit{coverage}}, \tag{7}$$

where

$$\beta_a = \sum_{u\in\mathcal{U}:\, a(u)=a} \pi_u, \qquad \alpha_a = \pi_a + \beta_a, \qquad \kappa_{\text{cov}} = 1 + L_H(2L_s + L_t).$$

The proof is given in Appendix C. Here, β_a is the total weight of unseen levels assigned to calibration level a, while $\alpha_a = \pi_a + \beta_a$ includes both the weight of a itself and those assigned unseen levels. The coverage quantity $\rho(\mathcal{A})$ is the weighted distance from unseen levels to their nearest calibration levels.

The weights α_a and β_a depend on how much target mass is assigned to calibration level a. Specifically, β_a sums the deployment weights of unseen levels whose nearest calibration level is a, while $\alpha_a = \pi_a + \beta_a$ also includes the weight at a itself. Calibration levels associated with more target mass therefore receive larger weights in the prediction and discrepancy terms. In contrast, $\rho(\mathcal{A})$ depends only on the placement of the calibration levels and is independent of ϕ. Thus, adaptation can reduce the prediction and discrepancy terms, but not the coverage term.

5 CaDRA: From the Robust Bound to the Empirical Objective

Theorem 2 decomposes the robust target risk into calibration-level prediction error, source–target discrepancy, and a coverage term. The coverage term is fixed by the calibration design, while the first two terms in Eq. (7) depend on ϕ. We now construct empirical surrogates for the two adaptation-dependent terms and obtain the training objective.

5.1 Empirical Surrogate

For each $a \in \mathcal{A}$, let $\mathcal{I}_t(a)$ and $\mathcal{I}_s(a)$ denote the indices of target and source samples at level a, with $n_t^a = |\mathcal{I}_t(a)|$ and $n_s^a = |\mathcal{I}_s(a)|$. The target prediction risk is estimated by the empirical calibration loss

$$\widehat{R}_a(\phi) = \frac{1}{n_t^a} \sum_{i \in \mathcal{I}_t(a)} \left| H_s(T_\phi(E_s(x_i^t), m_i^t)) - a \right|.$$

We next construct an empirical surrogate for the latent discrepancy $\Delta_\phi(a) = W_1(Q_\phi^a, P_s^a)$. Let $\widehat{Q}_\phi^a$ and $\widehat{P}_s^a$ denote the empirical adapted target and source latent distributions at level a.

Since samples collected at the same physical level may still differ in operating state, we match source and target samples using their observed operating descriptors. For $i \in \mathcal{I}_t(a)$ and $j \in \mathcal{I}_s(a)$, let $D_{ij}^a = D(m_i^t, m_j^s)$ denote the operating-condition distance. Define the set of balanced couplings as

$$\Pi_a = \left\{ \Gamma \geq 0 : \Gamma \mathbf{1} = \frac{1}{n_t^a}\mathbf{1},\ \Gamma^\top \mathbf{1} = \frac{1}{n_s^a}\mathbf{1} \right\}.$$

We choose the operating-aware coupling

$$\bar{\Gamma}_a \in \arg\min_{\Gamma \in \Pi_a} \sum_{i,j} \Gamma_{ij} D_{ij}^a. \tag{8}$$

The numerical solution of Eq. (8) is described in Appendix E. We use the operating descriptors to construct a balanced source–target coupling. The marginal constraints assign each source and target sample its empirical mass. Among all such couplings, Eq. (8) minimizes the total operating-condition distance. The resulting coupling is then used to measure source–target discrepancy in the latent space. The corresponding empirical alignment cost is

$$\widehat{A}_a(\phi) = \sum_{i,j} \bar{\Gamma}_{a,ij} \left\| T_\phi(E_s(x_i^t), m_i^t) - E_s(x_j^s) \right\|_2^2. \tag{9}$$

Because $\bar{\Gamma}_a$ is a valid coupling, $W_2^2(\widehat{Q}_\phi^a, \widehat{P}_s^a) \leq \widehat{A}_a(\phi)$ by the definition of W_2. Moreover, Jensen's inequality gives $W_1 \leq W_2$. Hence,

$$W_1(\widehat{Q}_\phi^a, \widehat{P}_s^a) \leq W_2(\widehat{Q}_\phi^a, \widehat{P}_s^a) \leq \sqrt{\widehat{A}_a(\phi)}. \tag{10}$$

Using the empirical calibration loss $\widehat{R}_a(\phi)$ and the alignment bound in Eq. (10), we define the surrogate to optimize

$$\widehat{\mathcal{S}}(\phi) = \sum_{a \in \mathcal{A}} \alpha_a \widehat{R}_a(\phi) + 2L_H \sum_{a \in \mathcal{A}} \beta_a \sqrt{\widehat{A}_a(\phi)}. \tag{11}$$

5.2 Training Objective

To express the surrogate in terms of the empirical alignment cost $\widehat{A}_a(\phi)$, we use Young's inequality $2L_H\sqrt{A} \le \lambda A + L_H^2/\lambda$ for any $\lambda > 0$, which gives

$$\widehat{\mathcal{S}}(\phi) \le \sum_{a\in\mathcal{A}} \alpha_a \widehat{R}_a(\phi) + \lambda \sum_{a\in\mathcal{A}} \beta_a \widehat{A}_a(\phi) + \frac{L_H^2}{\lambda} \sum_{a\in\mathcal{A}} \beta_a. \tag{12}$$

The final term is independent of ϕ. Defining

$$\mathcal{L}_{\text{cal}} = \sum_{a\in\mathcal{A}} \alpha_a \widehat{R}_a(\phi), \qquad \mathcal{L}_{\text{align}} = \sum_{a\in\mathcal{A}} \beta_a \widehat{A}_a(\phi),$$

gives the training objective

$$\min_{\phi} \ \mathcal{L}_{\text{cal}} + \lambda \mathcal{L}_{\text{align}}. \tag{13}$$

The coefficient λ is selected empirically.

We propose *CaDRA* (**Ca**libration-aware **D**istributionally **R**obust **A**daptation), defined by Eq. (13). CaDRA combines supervised prediction at the observed target levels with source–target alignment under the balanced operating-aware coupling. The calibration and alignment terms are weighted by α_a and β_a, respectively, as derived in Eq. (7). CaDRA uses target data only from the observed calibration levels during adaptation.

6 Experiments

Our experiments evaluate the performance and robustness of CaDRA. We first consider a smooth synthetic shift with gradual cross-level variation, and then introduce an abrupt shift with a sharp change across neighboring severity levels. We further vary the calibration set to assess robustness across different coverage patterns. Finally, we evaluate CaDRA on three real-world transfer tasks.

Our primary metric is mean absolute error (MAE) over the target condition range, obtained by averaging the MAE at each physical level equally across $\mathcal{C}$. Let $\mathcal{I}_t(c)$ denote the target test samples at level c. We compute

$$\text{Target MAE} = \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \frac{1}{|\mathcal{I}_t(c)|} \sum_{i\in\mathcal{I}_t(c)} \left| f_t^{\phi}(x_i^t, m_i^t) - c \right|.$$

6.1 Experimental Protocol

Target-data access. All experiments follow the same sparse-target protocol. Source data cover the full set $\mathcal{C}$, while target data are available only at calibration levels $\mathcal{A}$. Samples from unseen target levels $\mathcal{U}$ are excluded from training, checkpoint selection, and hyperparameter tuning.

Baselines. We compare against a no-adaptation baseline, denoted *Source Only*, which applies the pretrained source predictor directly to the target domain without target adaptation. We further compare with representative domain adaptation baselines: CORAL (Sun & Saenko, 2016), DANN (Ganin et al., 2016), RSD (Chen et al., 2021), and COD (Yang et al., 2024). We do not include LGOT (Chen et al., 2026), as it requires unlabeled target data beyond the observed calibration levels, which are unavailable in our setting. We also found no official public implementation of LGOT (Chen et al., 2026) at the time of submission.

Unless stated otherwise, all learned methods use the same frozen source encoder, prediction head, and target adapter, and receive the same target calibration data. All reported results are averaged over 5 random seeds. Implementation details and baseline-specific adaptations are provided in Appendix G.

6.2 Controlled Synthetic Experiments

The synthetic experiments evaluate two aspects of CaDRA under controlled conditions. We first compare performance under a smooth shift with gradual cross-level variation and an abrupt shift

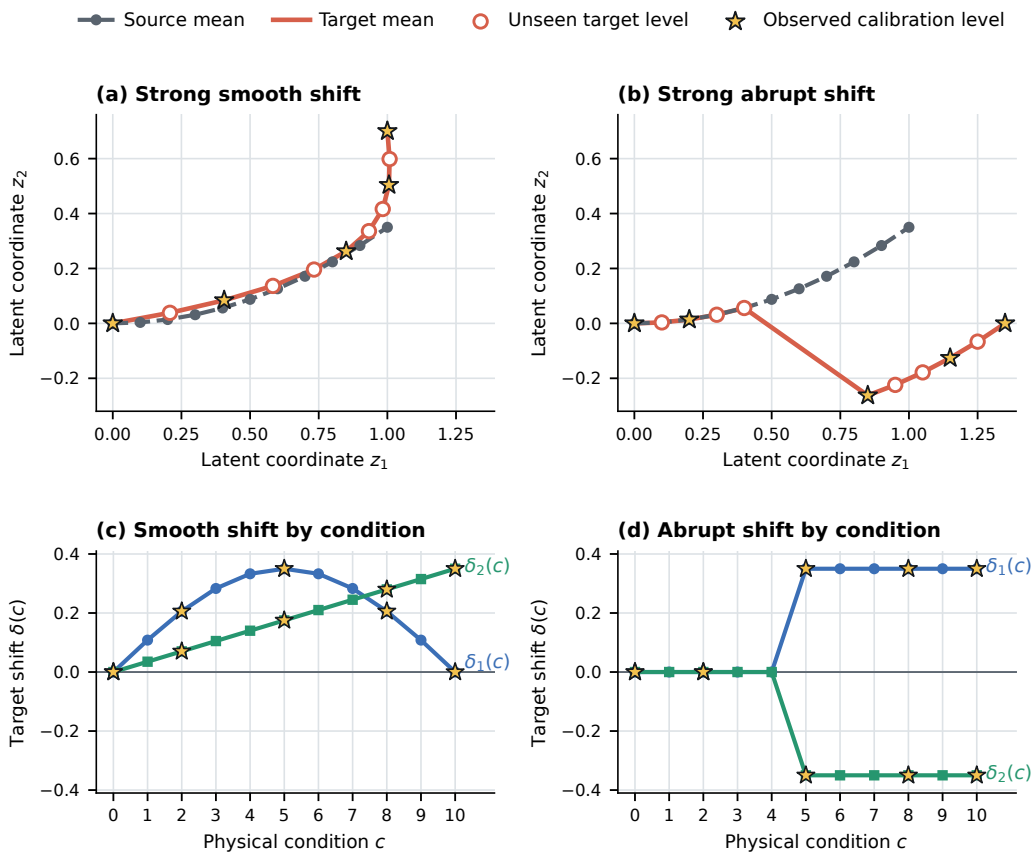


Figure 1: Geometry of the smooth and abrupt synthetic shifts. The top row shows the source and target conditional means, and the bottom row shows the corresponding shift components across physical conditions. Stars indicate observed target calibration levels.

with a sharp local change. We then vary the calibration set to assess robustness across different coverage patterns.

Smooth and Abrupt Domain Shifts: We construct two synthetic domain shifts with different degrees of cross-level variation. Let $c \in \{0, 1, \ldots, 10\}$ denote severity and sample the observed operating variable $m \sim \mathrm{Unif}[-1, 1]$. The source latent distribution is

$$Z_s \mid c, m \sim \mathcal{N}\left(\begin{bmatrix} c/10 + 0.2m \\ 0.35(c/10)^2 + 0.1m \end{bmatrix}, 0.05^2 I \right).$$

The target distribution has the same covariance, while its mean is shifted by either

$$\delta_{\mathrm{smooth}}(c) = 0.35 \begin{bmatrix} \sin(\pi c/10) \\ c/10 \end{bmatrix}, \qquad \delta_{\mathrm{abrupt}}(c) = \begin{cases} (0,0)^\top, & c < 5, \\ (0.35, -0.35)^\top, & c \geq 5. \end{cases}$$

The smooth shift varies gradually across severity levels, whereas the abrupt shift introduces a sharp change between neighboring levels. Figure 1 visualizes the two shifts.

Table 1 reports results under a fixed calibration set. All adaptation methods improve substantially over Source Only under both shifts. Under the smooth shift, the competing adaptation methods perform similarly, while CaDRA achieves the lowest error. Under the abrupt shift, the difference between CaDRA and the other adaptation methods increases, showing a larger advantage when the target distributions vary more sharply across neighboring severity levels.

Table 1: Synthetic Target MAE under smooth and abrupt shifts.

Shift	Source Only	CORAL	DANN	RSD	COD	CaDRA (Ours)
Smooth	1.961 ± 0.090	0.363 ± 0.013	0.368 ± 0.018	0.363 ± 0.016	0.362 ± 0.014	$\mathbf{0.343 \pm 0.004}$
Abrupt	0.951 ± 0.080	0.449 ± 0.014	0.446 ± 0.036	0.441 ± 0.036	0.456 ± 0.025	$\mathbf{0.383 \pm 0.007}$

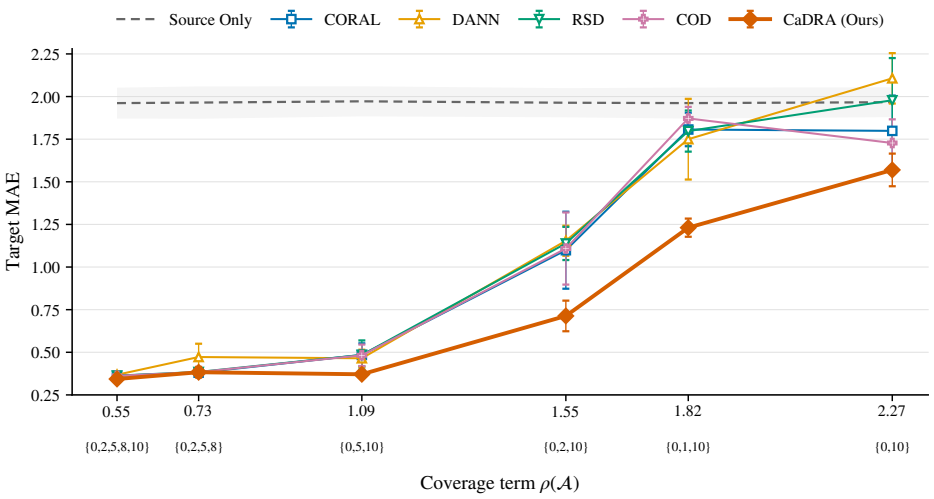


Figure 2: Synthetic performance across calibration designs. Target MAE is shown against the coverage term $\rho(\mathcal{A})$ for six calibration sets. Markers and error bars show the mean and standard deviation across random seeds.

Sensitivity to Calibration Design: We next vary the calibration set while keeping the underlying source–target shift fixed. The resulting designs differ in both the placement and the number of observed calibration levels, and therefore also in the coverage $\rho(\mathcal{A})$. Figure 2 reports Target MAE across these calibration designs. CaDRA achieves the lowest mean Target MAE in all six settings, indicating consistent performance across substantially different sparse-calibration patterns. Across these designs, larger $\rho(\mathcal{A})$ generally corresponds to more difficult adaptation, consistent with the role of the coverage term in Theorem 2.

6.3 Real-World Transfer Experiments

6.3.1 Datasets and Transfer Tasks

For all three datasets, the exact calibration sets and corresponding $\rho(\mathcal{A})$ values are provided in Appendix G.

MotorDA: cross-machine transfer. We use *MotorDA*, an anonymized name for a proprietary multimodal motor dataset collected for this study.[1] The dataset contains vibration, stator-current, acoustic, and magnetic-flux measurements from two physical motors under controlled rotor imbalance. We transfer from Motor 1 to Motor 2 and evaluate two calibration regimes with relatively better and worse coverage, referred to as *good* and *poor* coverage. Speed and load are used as the operating descriptors for matching.

CWRU: cross-sensor transfer. For CWRU (Case Western Reserve University Bearing Data Center, n.d.), we treat the drive-end and fan-end accelerometers as the source and target domains, respectively, and consider outer-race defects measured under shared operating conditions. This setting tests transfer across sensor locations while keeping the underlying machine and fault type fixed. We

[1]The name *MotorDA* is used for anonymity during double-blind review. Public release is planned following the patent review.

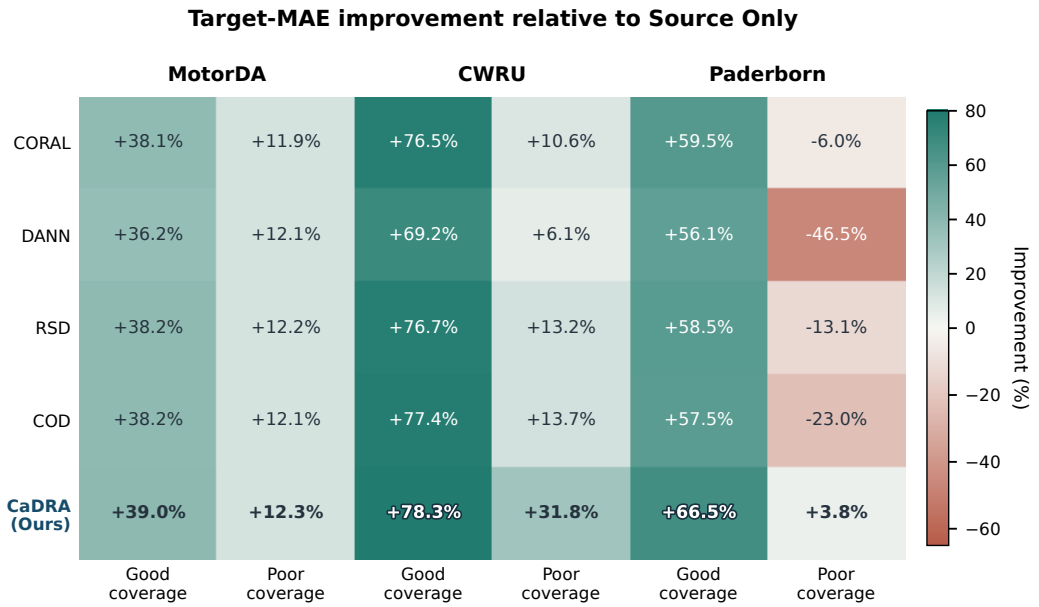


Figure 3: Percentage improvement in Target MAE over Source Only on the three real-world datasets. Bold entries indicate the lowest mean Target MAE in each dataset–coverage setting.

evaluate both good- and poor-coverage calibration regimes, with load and speed used as the operating descriptors for matching.

Paderborn: transfer across damage-generation processes. For Paderborn (Lessmeier et al., 2016), we use bearings with controlled outer-race damage as the source domain and bearings degraded through accelerated lifetime experiments as the target domain. This setting evaluates transfer between two different damage-generation processes for the same fault type. We evaluate both good- and poor-coverage calibration regimes under shared operating conditions, using speed, torque, and radial force as the operating descriptors for matching.

6.3.2 Transfer Performance

Figure 3 reports the percentage improvement in Target MAE over Source Only. The three datasets cover distinct transfer settings: cross-machine transfer on MotorDA, cross-sensor transfer on CWRU, and cross-process transfer on Paderborn. CaDRA achieves the lowest mean Target MAE in all six dataset–coverage settings. The same adaptation objective is used in all three cases, showing that CaDRA does not rely on a particular source of domain shift.

Appendix F further examines the calibration and alignment losses and the effect of balanced operating-aware matching.

7 Conclusion

We studied domain adaptation with sparse target calibration, where labeled target data cover only part of the target label space. Our analysis shows that the observed target distributions do not uniquely determine the target distributions at unseen levels, motivating a distributionally robust formulation and the calibration-and-alignment objective used in CaDRA. The bound shows that prediction and source–target discrepancy can be improved through adaptation, while the remaining coverage term is determined by the observed target levels. Experiments on controlled synthetic shifts and three real-world transfer tasks show consistent performance across different calibration regimes.

REFERENCES

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(1–2): 151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936. URL https://doi.org/10.1287/moor.2018.0936.

Case Western Reserve University Bearing Data Center. 12k drive end bearing fault data. Case School of Engineering, n.d. URL https://engineering.case.edu/bearingdatacenter/12k-drive-end-bearing-fault-data. Accessed 2026-08-22.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018. URL https://jmlr.org/papers/v19/17-295.html.

Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1749–1759. PMLR, 2021. URL https://proceedings.mlr.press/v139/chen21u.html.

Zi-Ying Chen, Chuan-Xian Ren, and Hong Yan. Label-guided optimal transport for domain adaptation regression. *Pattern Recognition*, 179:113697, November 2026. doi: 10.1016/j.patcog.2026.113697. URL https://doi.org/10.1016/j.patcog.2026.113697.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023. doi: 10.1287/moor.2022.1275. URL https://doi.org/10.1287/moor.2022.1275.

Pin Jiang, Aming Wu, Yahong Han, Yunfeng Shao, Meiyu Qi, and Bingshuai Li. Bidirectional adversarial training for semi-supervised domain adaptation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 934–940. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/130. URL https://doi.org/10.24963/ijcai.2020/130.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, INFORMS TutORials in Operations Research, pp. 130–166. INFORMS, 2019. doi: 10.1287/educ.2019.0198. URL https://doi.org/10.1287/educ.2019.0198.

Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. *PHM Society European Conference*, 3(1), 2016. doi: 10.36001/phme.2016.v3i1.1577. URL https://doi.org/10.36001/phme.2016.v3i1.1577.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105. PMLR, 2015. URL https://proceedings.mlr.press/v37/long15.html.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018. doi: 10.1007/s10107-017-1172-1. URL https://doi.org/10.1007/s10107-017-1172-1.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8049–8057. IEEE, October 2019. doi: 10.1109/ICCV.2019.00814. URL https://doi.org/10.1109/ICCV.2019.00814.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pp. 443–450. Springer International Publishing, 2016. ISBN 978-3-319-49409-8. doi: 10.1007/978-3-319-49409-8_35. URL https://doi.org/10.1007/978-3-319-49409-8_35.

Takeshi Teshima, Issei Sato, and Masashi Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9458–9469. PMLR, 2020. URL https://proceedings.mlr.press/v119/teshima20a.html.

Hao-Ran Yang, Chuan-Xian Ren, and You-Wei Luo. COD: Learning conditional invariant representation for domain adaptation regression. In *Computer Vision – ECCV 2024*, pp. 108–125. Springer Nature Switzerland, 2024. ISBN 978-3-031-73116-7. doi: 10.1007/978-3-031-73116-7_7. URL https://doi.org/10.1007/978-3-031-73116-7_7.

Siyu Zhang, Lei Su, Jiefei Gu, Ke Li, Lang Zhou, and Michael Pecht. Rotating machinery fault detection and diagnosis based on deep domain adaptation: A survey. *Chinese Journal of Aeronautics*, 36(1):45–74, 2023. doi: 10.1016/j.cja.2021.10.006. URL https://doi.org/10.1016/j.cja.2021.10.006.

Zepeng Zhang and Olga Fink. Domain adaptive unfolded graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22714–22722, 2025. doi: 10.1609/aaai.v39i21.34431. URL https://doi.org/10.1609/aaai.v39i21.34431.

AI Use Statement

Generative AI tools were used to assist with language editing and paper organization. All mathematical claims, proofs, citations, experiments, and final text were checked and approved by the authors, who take responsibility for the final content.

Ethics Statement

The study uses industrial machine-sensor measurements and does not involve human subjects or personally identifiable information. Proprietary data and company-sensitive information will be handled according to applicable organizational policies.

Reproducibility Statement

The main paper specifies the sparse-calibration protocol, adaptation objective, and theoretical assumptions. The appendix provides complete proofs, dataset protocols, model architectures, training details, baseline implementations, and the synthetic data generation process. Code and all non-proprietary experimental configurations will be released with the submission. The proprietary MotorDA recordings are subject to company release restrictions. All public-data experiments use publicly available datasets.

Throughout the appendix, the latent space is $\mathcal{Z} = \mathbb{R}^d$ with the Euclidean metric, and all latent distributions are assumed to have finite first moments.

A Regularity of Conditional Discrepancy

Recall that $\Delta_\phi(c) = W_1(Q_\phi^c, P_s^c)$.

Lemma 1 (Conditional discrepancy regularity)**.** *Suppose that, for all $c, c' \in \mathcal{C}$, $W_1(P_s^c, P_s^{c'}) \leq L_s|c - c'|$ and $W_1(Q_\phi^c, Q_\phi^{c'}) \leq L_t|c - c'|$. Then*

$$|\Delta_\phi(c) - \Delta_\phi(c')| \leq (L_s + L_t)|c - c'|.$$

Proof. By the triangle inequality for W_1,

$$\begin{aligned} |\Delta_\phi(c) - \Delta_\phi(c')| &= \left| W_1(Q_\phi^c, P_s^c) - W_1(Q_\phi^{c'}, P_s^{c'}) \right| \\ &\leq W_1(Q_\phi^c, Q_\phi^{c'}) + W_1(P_s^c, P_s^{c'}) \\ &\leq (L_s + L_t)|c - c'|. \end{aligned}$$

□

B Risk Transport Lemma

Lemma 2 (Wasserstein transport of regression risk)**.** *Suppose that H_s is L_H-Lipschitz. Then, for any latent distributions P, Q and any physical level c,*

$$|\mathbb{E}_{Z\sim Q}|H_s(Z) - c| - \mathbb{E}_{Z\sim P}|H_s(Z) - c|| \leq L_H W_1(Q, P).$$

Proof. Let $g_c(z) = |H_s(z) - c|$. For any z, z',

$$|g_c(z) - g_c(z')| \leq |H_s(z) - H_s(z')| \leq L_H \|z - z'\|_2.$$

Thus g_c is L_H-Lipschitz. The result follows from the Kantorovich–Rubinstein dual representation of W_1. □

C Proof of the Full Target Robust-Risk Bound

The proof proceeds in two steps. We first bound the robust risk at each unseen level by quantities evaluated at its nearest calibration level. We then aggregate these bounds over all unseen levels.

The key idea is to connect an unseen level u to its nearest calibration level $a = a(u)$ through the source domain. We first compare the robust risk at u with the source risk at u, then move the source risk from u to a, and finally compare the source and target risks at a.

Lemma 3 (Unseen-condition robust risk)**.** *Let $u \in \mathcal{U}$ and $a = a(u) \in \mathcal{A}$. Under the assumptions in Section 4,*

$$\sup_{Q\in\mathbb{B}_\phi(u)} \mathbb{E}_{Z\sim Q}|H_s(Z) - u| \leq \mathcal{R}_t(a;\phi) + 2L_H\Delta_\phi(a) + \kappa_{\mathrm{cov}}|u - a|,$$

where $\kappa_{\mathrm{cov}} = 1 + L_H(2L_s + L_t)$.

Proof. Define the source risk at level c by $\mathcal{R}_s(c) = \mathbb{E}_{Z\sim P_s^c}|H_s(Z) - c|$.

For any $Q \in \mathbb{B}_\phi(u)$, the risk-transport lemma and the definition of $\mathbb{B}_\phi(u)$ give

$$\begin{aligned} \mathbb{E}_{Z\sim Q}|H_s(Z) - u| &\leq \mathcal{R}_s(u) + L_H W_1(Q, P_s^u) \\ &\leq \mathcal{R}_s(u) + L_H\Delta_\phi(a) + L_H(L_s + L_t)|u - a|. \end{aligned}$$

Applying the risk-transport lemma to P_s^u and P_s^a, together with the source cross-level regularity, yields

$$\mathcal{R}_s(u) \leq \mathbb{E}_{Z\sim P_s^a}|H_s(Z) - u| + L_H L_s|u - a|.$$

Since $|H_s(Z) - u| \leq |H_s(Z) - a| + |u - a|$,

$$\mathcal{R}_s(u) \leq \mathcal{R}_s(a) + (1 + L_H L_s)|u - a|.$$

Moreover, $\mathcal{R}_t(a;\phi) = \mathbb{E}_{Z\sim Q_\phi^a}|H_s(Z) - a|$ and $\Delta_\phi(a) = W_1(Q_\phi^a, P_s^a)$. The risk-transport lemma therefore gives

$$\mathcal{R}_s(a) \le \mathcal{R}_t(a;\phi) + L_H\Delta_\phi(a).$$

Combining the inequalities,

$$\mathbb{E}_{Z\sim Q}|H_s(Z) - u| \le \mathcal{R}_t(a;\phi) + 2L_H\Delta_\phi(a) + \big[1 + L_H(2L_s + L_t)\big]|u - a|.$$

Taking the supremum over $Q \in \mathbb{B}_\phi(u)$ proves the result. □

Proof of Theorem 2. Since $Q_\phi^u \in \mathbb{B}_\phi(u)$ for every $u \in \mathcal{U}$, the true target risk at each unseen level is bounded by its worst-case counterpart. The calibration-level terms are identical in $\mathcal{R}_t^{\mathrm{all}}$ and $\mathcal{R}_t^{\mathrm{robust}}$, hence $\mathcal{R}_t^{\mathrm{all}}(\phi) \le \mathcal{R}_t^{\mathrm{robust}}(\phi)$.

Applying the preceding lemma to every $u \in \mathcal{U}$ gives

$$\begin{aligned}\mathcal{R}_t^{\mathrm{robust}}(\phi) \le \sum_{a\in\mathcal{A}} \pi_a\mathcal{R}_t(a;\phi) + \sum_{u\in\mathcal{U}} \pi_u\mathcal{R}_t(a(u);\phi)\\ + 2L_H\sum_{u\in\mathcal{U}} \pi_u\Delta_\phi(a(u)) + \kappa_{\mathrm{cov}}\sum_{u\in\mathcal{U}} \pi_u|u - a(u)|.\end{aligned}$$

Grouping the unseen levels by their nearest calibration level and using $\beta_a = \sum_{u\in\mathcal{U}:,a(u)=a} \pi_u$, $\alpha_a = \pi_a + \beta_a$, and $\rho(\mathcal{A}) = \sum_{u\in\mathcal{U}} \pi_u|u - a(u)|$, we obtain

$$\mathcal{R}_t^{\mathrm{robust}}(\phi) \le \sum_{a\in\mathcal{A}} \alpha_a\mathcal{R}_t(a;\phi) + 2L_H\sum_{a\in\mathcal{A}} \beta_a\Delta_\phi(a) + \kappa_{\mathrm{cov}}\rho(\mathcal{A}),$$

which is Eq. (7). □

D Proof of the Calibration-Coverage Lower Bound

Proof of Theorem 1. Let $d_{\mathcal{A}}(c) = \min_{a\in\mathcal{A}}|c - a|$, and let $e \in \mathbb{R}^d$ be any unit vector. For $\sigma \in \{-1, +1\}$, define

$$Q_\sigma^c = \delta_{\sigma L_t d_{\mathcal{A}}(c)e}, \qquad c \in \mathcal{C}.$$

Since $d_{\mathcal{A}}$ is 1-Lipschitz,

$$W_1(Q_\sigma^c, Q_\sigma^{c'}) = L_t|d_{\mathcal{A}}(c) - d_{\mathcal{A}}(c')| \le L_t|c - c'|,$$

so both families belong to $\mathfrak{Q}(L_t)$.

For every $a \in \mathcal{A}$, $d_{\mathcal{A}}(a) = 0$, and hence $Q_+^a = Q_-^a = \delta_0$. Thus, even exact knowledge of all calibration-level target conditionals cannot distinguish the two families. The source distributions can be chosen identically under both constructions and therefore provide no additional information.

Any estimator based on the available information must consequently return the same $\widehat{Q}^u$ under the two families. For each $u \in \mathcal{U}$, the triangle inequality gives

$$W_1(\widehat{Q}^u, Q_+^u) + W_1(\widehat{Q}^u, Q_-^u) \ge W_1(Q_+^u, Q_-^u) = 2L_t d_{\mathcal{A}}(u).$$

Multiplying by π_u and summing over $u \in \mathcal{U}$ yields

$$\sum_{u\in\mathcal{U}} \pi_u W_1(\widehat{Q}^u, Q_+^u) + \sum_{u\in\mathcal{U}} \pi_u W_1(\widehat{Q}^u, Q_-^u) \ge 2L_t\rho(\mathcal{A}).$$

Therefore, at least one of the two families incurs weighted estimation error at least $L_t\rho(\mathcal{A})$. Taking the supremum over $\mathbf{Q} \in \mathfrak{Q}(L_t)$ and then the infimum over $\widehat{\mathbf{Q}}$ proves Eq. (4). □

E Balanced Matching and the Alignment Bound

At a fixed calibration level a, let the empirical target and source measures be

$$\widehat{\nu}_t^a = \frac{1}{n_t^a} \sum_{i \in \mathcal{I}_t(a)} \delta_{(x_i^t, m_i^t)}, \qquad \widehat{\nu}_s^a = \frac{1}{n_s^a} \sum_{j \in \mathcal{I}_s(a)} \delta_{(x_j^s, m_j^s)}.$$

The coupling $\bar{\Gamma}_a$ defined in Eq. (8) belongs to Π_a, and is therefore a valid coupling of $\widehat{\nu}_t^a$ and $\widehat{\nu}_s^a$. Pushing this coupling forward through

$$(x_t, m_t) \mapsto T_\phi(E_s(x_t), m_t), \qquad x_s \mapsto E_s(x_s)$$

gives a coupling between $\widehat{Q}_\phi^a$ and $\widehat{P}_s^a$. Hence,

$$W_2^2(\widehat{Q}_\phi^a, \widehat{P}_s^a) \le \sum_{i,j} \bar{\Gamma}_{a,ij} \left\| T_\phi(E_s(x_i^t), m_i^t) - E_s(x_j^s) \right\|_2^2 = \widehat{A}_a(\phi).$$

Since $W_1 \le W_2$, Eq. (10) follows.

In practice, we compute an entropically regularized coupling

$$\Gamma_a^\varepsilon = \arg\min_{\Gamma \in \Pi_a} \left\{ \sum_{i,j} \Gamma_{ij} D_{ij}^a + \varepsilon_{\mathrm{OT}} \sum_{i,j} \Gamma_{ij} (\log \Gamma_{ij} - 1) \right\}.$$

Because $\Gamma_a^\varepsilon \in \Pi_a$, the same alignment bound applies when Γ_a^ε is used in place of $\bar{\Gamma}_a$. The value of $\varepsilon_{\mathrm{OT}}$ and the numerical solver are given in Appendix G.

F MotorDA Dataset and Ablation

F.1 Dataset Collection and Composition

MotorDA was collected in house on a motor–fan test rig using two physical motors. Rotor imbalance is induced by attaching controlled masses to a fan blade. The common evaluation levels are $\{0, 5, 7.5, 10, 12.5\}$ g, where 0 g denotes healthy operation.

Motor 1 contains 1,323 recordings and 15,467 nonoverlapping two-second segments, while Motor 2 contains 6,693 recordings and 64,297 segments. At the common severity levels, the recording counts are $(276, 275, 251, 271, 250)$ for Motor 1 and $(1080, 1079, 1296, 1080, 1079)$ for Motor 2.

Synchronized acquisition includes three-phase voltage and current, three-axis vibration, magnetic flux, acoustic signals, and tachometer speed. The experiments span five air-duct closure settings $(0, 25, 50, 75, 100)\%$ and multiple speed profiles. Motor 1 contains constant, fixed-profile, and random-speed runs, whereas Motor 2 contains steady-speed, ramp-up, ramp-down, and smooth-random runs. The experiments in this paper use vibration, stator current, acoustic, and magnetic-flux measurements. Additional preprocessing and operating-descriptor details are provided in Appendix G.

F.2 Hyperparameter Selection

For each dataset, we select the alignment coefficient λ using validation data from the observed calibration levels only. The same search space is used across datasets, and target samples from unseen levels are never used for hyperparameter selection. The selected value is $\lambda = 0.1$ for all reported experiments.

F.3 Objective and Matching Ablation

We evaluate the objective and matching components on the MotorDA good-coverage setting, with $\mathcal{A} = \{0, 7.5, 12.5\}$ g. The first two variants isolate the calibration and alignment losses. The remaining variants use both losses and differ only in the source–target matching scheme. Independent nearest matching selects source samples by operating-descriptor distance without a balance constraint;

Table 2: MotorDA objective and matching ablation with $\mathcal{A} = \{0, 7.5, 12.5\}$ g. Alignment-only and CaDRA use balanced operating-aware matching. Bold indicates the lowest mean Target MAE.

Variant	Calibration	Alignment	Balanced	Operating-aware	Target MAE ↓
Calibration only	✓	–	–	–	1.670 ± 0.026
Alignment only	–	✓	✓	✓	1.777 ± 0.036
Calibration + independent nearest	✓	✓	–	✓	1.702 ± 0.059
Calibration + balanced uniform	✓	✓	✓	–	1.683 ± 0.031
CaDRA (full)	✓	✓	✓	✓	$\mathbf{1.660 \pm 0.023}$

balanced uniform matching preserves the empirical marginals without using operating descriptors; CaDRA uses balanced operating-aware matching.

CaDRA achieves the lowest mean Target MAE. Among the variants using both losses, the mean decreases from 1.702 with independent nearest matching to 1.683 with balanced uniform matching and 1.660 with balanced operating-aware matching.

G Implementation Details

G.1 Calibration Sets and Coverage

For MotorDA, the evaluation levels are $\mathcal{C} = \{0, 5, 7.5, 10, 12.5\}$ g. The good-coverage calibration set is $\mathcal{A} = \{0, 7.5, 12.5\}$ g, with $\rho(\mathcal{A}) = 1.0$ g, while the poor-coverage set is $\mathcal{A} = \{0, 5, 7.5\}$ g, with $\rho(\mathcal{A}) = 1.5$ g.

For CWRU, the evaluation levels are $\mathcal{C} = \{0, 0.007, 0.014, 0.021\}$ inches. The good-coverage calibration set is $\mathcal{A} = \{0, 0.007, 0.021\}$, with $\rho(\mathcal{A}) = 0.00175$ inches, while the poor-coverage set is $\mathcal{A} = \{0, 0.007\}$, with $\rho(\mathcal{A}) = 0.00525$ inches.

For Paderborn, the evaluation levels are $\mathcal{C} = \{0, 1, 2\}$. The good-coverage calibration set is $\mathcal{A} = \{0, 2\}$, with $\rho(\mathcal{A}) = 1/3$, while the poor-coverage set is $\mathcal{A} = \{2\}$, with $\rho(\mathcal{A}) = 1$.

G.2 Synthetic Experiments

We use severities $c \in \{0, 1, \ldots, 10\}$ and sample $M \sim \mathrm{Unif}[-1, 1]$. Source features follow

$$Z_s \mid c, m \sim \mathcal{N}\left(\begin{bmatrix} c/10 + 0.2m \\ 0.35(c/10)^2 + 0.1m \end{bmatrix}, 0.05^2 I \right).$$

The target distribution has the same covariance and adds either the smooth shift $0.35[\sin(\pi c/10), c/10]^\top$ or the abrupt shift $(0, 0)^\top$ for $c < 5$ and $(0.35, -0.35)^\top$ otherwise. Each seed uses 1,000 source samples and 1,000 target test samples per severity, together with 400 target adaptation samples per calibration level.

The source regressor is a two-hidden-layer MLP of width 48, and the target adapter is a two-layer residual MLP of width 32 conditioned on (Z, M). Training configurations are fixed before evaluation on the target test set, which is used only for final performance reporting.

G.3 MotorDA Protocol

Motor 1 is used as the source domain and Motor 2 as the target domain. All methods use the same fixed preprocessing and two-second signal segments for vibration, stator current, acoustic, and magnetic-flux measurements. Features are normalized using source-training statistics.

Operating descriptors. Each recording is associated with a seven-dimensional operating descriptor containing the 10th and 90th percentiles of measured shaft speed, their difference, the recorded load setting, and three indicators for the constant, fixed, and random speed families. The two speed percentiles are normalized by 1500 rpm, the speed range by 500 rpm, and the load percentage by 100. Matching uses Euclidean distance between the normalized descriptors.

Source model. The pretrained backbone processes modality-specific temporal tokens and aggregates them into a 48-dimensional representation. Source recordings are split into training, validation, and test sets with a 70/15/15 ratio. The source model is trained once on Motor 1, after which the encoder and linear regression head remain fixed throughout target adaptation.

Target adapter. The adapter is $T_\phi(z, m) = z + \Delta_\phi^{\mathrm{net}}(z, m)$, a 9,902-parameter FiLM residual network. Its representation branch uses LayerNorm–Linear–GELU–Linear layers at width 48, while a second branch maps the seven-dimensional operating descriptor to feature-wise scale and bias. We train the adapter with AdamW using learning rate 8×10^{-4}, weight decay 2×10^{-4}, batch size 64, at most 200 epochs, and patience 40.

Balanced matching. The operating-aware coupling is computed once before adaptation and kept fixed during training. We use Euclidean descriptor cost and entropic Sinkhorn regularization with $\varepsilon_{\mathrm{OT}} = 0.05$, at most 2,000 iterations, and marginal tolerance 10^{-7}.

G.4 CWRU Protocol

We use the 12-kHz Case Western Reserve University bearing dataset with outer-race defects centered at the 6-o'clock load-zone position. Drive-end vibration is treated as the source domain and fan-end vibration as the target domain. Healthy recordings are provided by files 97–100, while files 130–133, 197–200, and 234–237 correspond to defect diameters 0.007, 0.014, and 0.021 inches, respectively.

Normal recordings are decimated from 48 kHz to 12 kHz. Each recording is split into nonoverlapping 2,048-sample windows after assigning the first 60%, next 20%, and final 20% to training, validation, and test, respectively. Each standardized window is represented by 256 averaged log-magnitude spectral bands.

Load and measured shaft speed form the operating descriptor. At each observed calibration level, we compute a balanced source–target coupling using load–speed descriptor distance while preserving uniform empirical marginals. Only target samples from the observed calibration levels are used during adaptation.

G.5 Paderborn Protocol

We use the Paderborn bearing dataset with outer-race damage. Artificially damaged bearings KA01, KA03, KA05, KA06, KA07, KA08, and KA09 form the source domain, while accelerated-lifetime bearings KA04, KA15, KA16, KA22, and KA30 form the target domain. Distinct healthy bearings are assigned to the source domain (K002–K004) and target domain (K001, K005, and K006). The values $0, 1, 2$ encode ordered damage extents and are used as the numerical severity scale in this experiment.

Each four-second measurement is divided into four nonoverlapping one-second windows. For vibration and each of the two phase-current channels, we standardize each window, compute its log-magnitude spectrum, and average it into 256 frequency bands. Concatenating the three channels gives a 768-dimensional representation. Measurement repetitions 1–12, 13–16, and 17–20 are used for training, validation, and test, respectively, so windows from the same measurement do not cross data splits.

Shaft speed, torque, and radial force form the operating descriptor. At each observed calibration level, we compute a balanced source–target coupling using operating-descriptor distance while preserving uniform empirical marginals. Only target samples from the observed calibration levels are used during adaptation. One unreadable KA08 measurement is omitted, leaving 3,196 source and 2,560 target windows.

H Baseline Implementation Details

All learned baselines use the same frozen source encoder, prediction head, target adapter, preprocessing, and target calibration data as CaDRA.

CORAL. We apply covariance alignment between source and target latent representations at the observed calibration levels, together with the target regression loss. The CORAL coefficient is fixed to 0.25 across all experiments.

DANN. We use a domain discriminator with gradient reversal to align source and target latent representations. Only target samples from the observed calibration levels are used during adaptation. The adversarial coefficient is fixed to 0.25 across all experiments.

COD. We use the COD objective, which combines conditional source–target alignment with marginal covariance alignment. Both terms are computed using target samples from the observed calibration levels.

RSD. We use the RSD subspace-alignment objective together with its basis-mismatch term. For the 48-dimensional latent representation, we use rank-32 subspaces to avoid the degenerate full-rank case. The adaptation coefficient is fixed to 0.001, and the basis-mismatch coefficient is fixed to 0.1.