Are Domain Generalization Benchmarks with Accuracy on the Line Misspecified?

Anonymous authors
Paper under double-blind review

Abstract

Spurious correlations are unstable statistical associations that hinder robust decision-making. Conventional wisdom suggests that models relying on such correlations will fail to generalize out-of-distribution (OOD), particularly under strong distribution shifts. However, a growing body of empirical evidence challenges this view, as naive empirical risk minimizers often achieve the best OOD accuracy across popular OOD generalization benchmarks. In light of these counterintuitive results, we propose a different perspective: *many widely used benchmarks for assessing the impact of spurious correlations on OOD generalization are misspecified.* Specifically, they fail to include shifts in spurious correlations that meaningfully degrade OOD generalization, making them unsuitable for evaluating the benefits of removing such correlations. We establish sufficient—and in some cases necessary—conditions under which a distribution shift can reliably assess a model's reliance on spurious correlations. Crucially, under these conditions, we provably should not observe a strong positive correlation between in-distribution and out-of-distribution accuracy—often referred to as accuracy on the line. Yet, when we examine state-of-the-art OOD generalization benchmarks, we find that most exhibit accuracy on the line, suggesting they do not effectively assess robustness to spurious correlations. Our findings expose a limitation in evaluating algorithms for domain generalization, i.e., learning predictors that do not rely on spurious correlations. Our results highlight the need to rethink how we assess robustness to spurious correlations.

1 Introduction

Domain generalization aims to develop predictors that generalize to arbitrarily new and potentially *worst-case* unobserved distributions (Arjovsky et al., 2019; Rosenfeld et al., 2020). Spurious correlations, also referred to as shortcuts, are coincidental statistical associations between features and labels in training data that fail to generalize beyond the training distribution, hindering domain generalization (Nagarajan et al., 2020; Geirhos et al., 2020; Makar et al., 2022). Thus, many algorithms for domain generalization have been focused on learning predictors that ignore these unreliable patterns—i.e., *invariance* or *feature disentanglement* methods (Arjovsky et al., 2019; Krueger et al., 2021; Wang et al., 2019; Parascandolo et al., 2020; Creager et al., 2021; Ahuja et al., 2021; Shi et al., 2021; Zhou et al., 2022; Wang et al., 2022b; Li et al., 2022; Salaudeen & Koyejo, 2024). Zhou et al. (2022); Wang et al. (2022b) provide a more comprehensive survey on domain generalization. However, empirical evidence suggests that standard empirical risk minimization (ERM), which may leverage spurious correlations (Xiao et al., 2020), often achieves the highest out-of-distribution (OOD) accuracy on widely used benchmarks (Gulrajani & Lopez-Paz, 2020a; Yao et al., 2022; Gagnon-Audet et al., 2022; Yang et al., 2023). Additionally, Taori et al. (2020); Miller et al. (2020; 2021); Wenzel et al. (2022); Baek et al. (2022); Saxena et al. (2024); Nastl & Hardt (2024) demonstrate a strong correlation between in- and out-of-distribution accuracy across several state-of-the-art benchmarks, suggesting that better in-distribution (ID) performance generally predicts better out-of-distribution performance. Nastl & Hardt (2024) further shows that a model that uses all available features generally results in better OOD performance than a model that uses a select subset of *causal features* for popular distribution shift tabular datasets (Gardner et al., 2023). This observation is seemingly counter to the thesis that there can be statistical patterns in training data that can improve ID performance but worsen OOD performance (Geirhos et al., 2020; Makar et al., 2022). These findings raise a critical question: *Does a better in-distribution classifier imply a better out-of-distribution classifier, challenging the necessity of targeted algorithms for domain generalization?*

We propose and provide evidence for an alternative explanation of ERM's success in domain generalization. Drawing on the concept of underspecification in modern machine learning pipelines (D'Amour et al., 2022), we argue that ERM's apparent superiority may stem from the (mis/under)specification of popular domain generalization benchmarks.

Specifically, we consider the setting where correlations exist in training data and a subset (spurious) shift at test time. We investigate the types of shifts under which predictions from a classifier that relies on these spurious correlations generalize worse out-of-distribution than a classier that only uses the static (domain-general) correlations, i.e., naive ERM is insufficient. We call these types of shifts *well-specified*. We argue that when the best in-distribution classifier is also the best out-of-distribution classifier on a benchmark, this benchmark does not represent the types of settings domain generalization is concerned with. The core of this work studies a benchmark's ability to distinguish between two types of predictors in this setting: one that ignores spurious correlations (domain-general or invariant) and another that leverages all available correlations (including spurious) that maximize in-distribution accuracy (domain-specific).

1.1 Our Contributions

- We establish a negative margin under distribution shift condition (or spurious correlation reversal) that characterizes well-specified domain generalization benchmarks—Theorem 1.
- We show that well-specified benchmarks typically lack a strong correlation between in and out-of-distribution accuracy for a diverse set of classifiers—such a strong correlation has been termed *accuracy on the line* (Miller et al., 2020)—Theorem 2 and Corollary 1). **When there is positive accuracy on the line, the dataset may be misspecified for evaluating domain generalization.**
- With over 40 total ID/OOD splits across 12 benchmarks, we show that many of the state-of-the-art benchmarks exhibit (*accuracy on the line*) and may be misspecified for domain generalization. Code*

While the theoretical and empirical results we present subsequently provide actionable insights for designing and using domain generalization benchmarks, we will also discuss their implications on benchmarking norms and practices, e.g., model selection, algorithmic fairness, causal representation learning, etc., in the context of both predictive and generative models.

2 Theoretical Analysis: (Mis)specification of Domain Generalization Benchmarks

Following previous work (Wang et al., 2019; Rosenfeld et al., 2020; Salaudeen & Koyejo, 2024) we define domain-general (dg) features $\mathcal{Z}_{\text{dg}} \subseteq \mathbb{R}^k$, where the optimal predictor that only uses domain-general features is desired. Conversely, spurious features (spu or domain-specific) $\mathcal{Z}_{\text{spu}} \subseteq \mathbb{R}^l$ contain additional domain-specific information that improves the prediction task in-distribution, but their use can degrade performance out-of-distribution. The observed features $\mathcal{X} \subseteq \mathbb{R}^d$ are a concatenation of $\mathcal{Z}_{\text{dg}}$ and $\mathcal{Z}_{\text{spu}}$, where $d = k + l$. We also define $\mathcal{Y} = \{\pm 1\}$. P's represent probability distributions over X, Y. Let $\mathcal{E}$, where $|\mathcal{E}| > 1$, denote the set of distributions of interest. $P \in \mathcal{E}$ implies that marginals without Z_{spu} are preserved.

We consider classifiers $f \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{Y}$ of the form $f(X) = Z_{\text{dg}}^T \text{w}_{\text{dg}} + Z_{\text{spu}}^T \text{w}_{\text{spu}}$ where $\text{w}_{\text{dg}} \in \mathbb{R}^k$ and $\text{w}_{\text{spu}} \in \mathbb{R}^l$. We note that our analysis generalizes to Z_{dg} and Z_{spu} from non-linear transformation when a final linear mapping is applied, e.g., kernel regressors or common deep neural networks; Rosenfeld et al. (2022a) demonstrate that differences in out-of-distribution performance are primarily due to differences in the learned linear classifier on learned (non-linear) representations. We revisit this point in our empirical findings.

Within $\mathcal{F}$, $\mathcal{F}_{\text{dg}} \subset \mathcal{F}$ comprises functions f_{dg} that only use domain-general features $f_{\text{dg}}(X) = f_{\text{dg}}([Z_{\text{dg}}; 0])$. Additionally, denote $f_{\text{X}} \in \mathcal{F} \backslash \mathcal{F}_{\text{dg}} := \mathcal{F}_{\text{X}}$, where f_{X} uses both domain-general and domain-specific features.

*`https://anonymous.4open.science/status/misspecified_DG_benchmarks-E5FD`.

The function $\ell(\cdot,\cdot) \mapsto \mathbb{R}$ denotes a loss function, and $\mathcal{R}^e(f) = \mathbb{E}_{P_e}[\ell(Y, f(X))]$ defines the expected loss for some function $f \in \mathcal{F}$. Additionally, we define the accuracy of $f \in \mathcal{F}$ on distribution P as

$$\mathrm{acc}_P(f) = \mathbb{E}_{(X,Y)\sim P}\big[\mathbf{1}(f(X)\cdot Y > 0)\big] = \Pr\big(f(X)\cdot Y > 0\big) \tag{1}$$

We first formally define spurious features and domain-general features (Definitions 1- 2). The relationship between spurious features and the label we want to predict is allowed to change across domains and can negatively impact out-of-distribution performance, while the relationship between domain-general features and the label is stable across domains. For instance, when predicting medical diagnoses from chest X-rays, the relationship between physiological features and diagnoses is expected to be stable from site to site (domain-general), while the relationship between site-specific markings on X-rays and diagnoses is unstable (spurious); models relying on site-specific markings fail out-of-distribution (Zech et al., 2018)

Definition 1 (Spurious Features)**.** *For all P_i, $P_j \in \mathcal{E}$,*

$$P_i(Y \mid Z_{spu}) \neq P_j(Y \mid Z_{spu}). \tag{2}$$

Definition 2 (Domain-General Features)**.** *For all P_i, $P_j \in \mathcal{E}$,*

$$P_i(Y \mid Z_{dg}) = P_j(Y \mid Z_{dg}). \tag{3}$$

We assume both types of features are informative about labels (Assumption 1) and are not redundant (Assumption 2). Clearly, if Assumption 1 does not hold, then the learning problem is misspecified; features are uncorrelated with labels. When Assumption 2 does not hold, spurious features are redundant and have no unique information about the target. Ahuja et al. (2021) study this setting (*Fully Informative Invariant Features (FIIF)*; we use 'domain-general' instead of 'invariant') and show that, under some conditions, models using spurious correlations can achieve equal OOD accuracy as the optimal invariant model. Hence, we focus on the partially informative domain-general features setting.

Assumption 1 (Informative Domain-General and Domain-Specific Features)**.** *For all observed training distributions P, $\mathbb{E}_P[Y \mid Z_{dg}] \neq \mathbb{E}_P[Y]$ and $\mathbb{E}_P[Y \mid Z_{spu}] \neq \mathbb{E}_P[Y]$.*

Assumption 2 (Non-Redundant Features)**.** *$Z_{spu} \not\perp\!\!\!\perp Y \mid Z_{dg}$ and $Z_{dg} \not\perp\!\!\!\perp Y \mid Z_{spu}$.*

Since we consider any feature whose inclusion decreases worst-case performance on $\mathcal{E}$ as spurious, the definition of spurious is strongly tied to the $\mathcal{E}$ 'worst-case' is with respect to. What is considered spurious for $\mathcal{E}' \neq \mathcal{E}$ may differ, even if $\mathcal{E}' \subset \mathcal{E}$. This observation implies that domain generalization cannot practically be divorced from domain expertise in defining $\mathcal{E}$. Clearly, too narrow of an $\mathcal{E}$ decreases expected robustness (potentially catastrophically), and too broad of an $\mathcal{E}$ may excessively and unnecessarily decrease overall utility.

We define two models: (i) the optimal domain-general model, which depends on $\mathcal{E}$ (Definition 3) and (ii) the optimal domain-specific model for a given $P \in E$ (Definition 4).

Definition 3 (Optimal Domain General Model $\mathbf{f}^{\mathcal{E}}_{\mathrm{dg}}$)**.** *Given a set of distributions of interest $\mathcal{E} = \{P_i(X,Y) : i = 1,\ldots\}$, $f^{\mathcal{E}}_{dg} = \mathrm{argmax}_{f\in\mathcal{F}_{dg}} \min_{P_i\in\mathcal{E}} acc_{P_i}(f)$. By construction, $f^{\mathcal{E}}_{dg} \in \mathcal{F}_{dg}$ does not use spurious features.*

Definition 4 (Optimal Domain Specific Model $\mathbf{f}^{\mathbf{P}}_{\mathrm{X}}$)**.** *Given a distribution P and $f^P_X = \mathrm{argmax}_{f\in\mathcal{F}} acc_P(f)$. By construction, $f^P_X \in \mathcal{F}_X$ uses spurious features.*

With informative and non-redundant features (Assumptions 1-2), Lemma 1 demonstrates that for a distribution $P \in \mathcal{E}$, the optimal domain-general and domain-specific models are different, and the optimal domain-specific model achieves a lower loss in-distribution than the optimal domain-general model.

Lemma 1 (Domain-General and Domain-Specific In-Distribution Error Gap)**.** *Assume Partially Informative Domain-General Features (Assumption 2) and strongly convex ℓ.*

$$\min_{f\in\mathcal{F}} \mathbb{E}_{(X,Y)\sim P}\big[\ell(f(X),Y)\big] < \min_{f\in\mathcal{F}_{dg}} \mathbb{E}_{(X,Y)\sim P}\big[\ell(f(X),Y)\big], \tag{4}$$

where $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ where $f(x) = \sigma((w)^\top x) = \sigma(w_{dg}^\top z_{dg} + w_{spu}^\top z_{spu})$, $f \in \mathcal{F}$. For $f \in \mathcal{F}_{dg}$, $f(x) = \sigma((w)^\top x) = \sigma(w_{dg}^\top z_{dg})$, where σ is the sigmoid function.

The proof of Lemma 1 is provided in Appendix A.1.

Given that the in-distribution risk minimizer and the domain-general model differ, we can now show that the domain-general model $f_{\text{dg}}^{\mathcal{E}}$ is only required for robustness under certain types of distribution shifts. Said differently, given some training and test distributions $P_{\text{ID}} \neq P_{\text{OOD}} \in \mathcal{E}$, respectively, the domain-general model $f_{\text{dg}}^{\mathcal{E}}$ does not always achieve a higher OOD accuracy than in-distribution risk minimizer $f_{\text{X}}^{P_{\text{ID}}}$ on P_{OOD}. Thus, given a P_{ID}, we derive sufficient conditions on P_{OOD} such that $f_{\text{dg}}^{\mathcal{E}}$ archives a higher OOD accuracy than $f_{\text{X}}^{P_{\text{ID}}}$. Such $P_{\text{ID}}, P_{\text{OOD}}$ ID/OOD splits make for 'well-specified' benchmarks, as outlined in the following.

Definition 5 (Well-Specified Domain Generalization Benchmark)**.** *Two ID/OOD splits, $P_{ID}, P_{OOD} \in \mathcal{E}$, are 'well-specified' if and only if*

$$acc_{P_{OOD}}(f_X^{P_{ID}}) < acc_{P_{OOD}}(f_X^{\mathcal{E}}), \tag{5}$$

where $f_{dg}^{\mathcal{E}}$, $f_X^{P_{ID}}$ are from Definitions 3 and 4, respectively.

Our results consider sub-Gaussian spurious features. We will consider an $M \in RR^{l\times l}$ that induces a shift such that $Z_{\text{spu}}^{\text{OOD}} = MZ_{\text{spu}}^{\text{ID}}$. We first identify shifts in spurious correlations that imply that a domain-general model achieves a higher out-of-distribution accuracy than a domain-specific predictor. Identifying such shifts is necessary to develop domain-generalization benchmarks where achieving the higher OOD accuracy meaningfully maps to learning domain-general predictors.

Theorem 1 (Condition for Well-Specified Domain Generalization Splits)**.** *Assume Z_{spu}^{ID} is sub-Gaussian with mean μ_{spu}, variance Σ_{spu}, and parameter κ. Define $M \in \mathbb{R}^{l\times l}$ such that $Z_{spu}^{OOD} = MZ_{spu}$ where in-distribution $Z_{spu}^{OOD} \sim P_{OOD}$ and out-of-distribution $Z_{spu}^{OOD} \sim P_{OOD}$. Additionally, denote w_{spu} as the weights learned by the optimal P_{ID} predictor ($f_X^{P_{ID}}$) for Z_{spu}^{ID}. Then, for any $\delta \in (0,1)$, if*

$$w_{spu}^\top M\mu_{spu} + \sqrt{2\kappa^2\, w_{spu}^\top M\Sigma_{spu}M^\top w_{spu}\,\log(1/\delta)} < 0, \tag{6}$$

with probability at least $1-\delta$ over Z_{spu}^{OOD}, we have

$$acc_{P_{OOD}}(f_X^{P_{ID}}) < acc_{P_{OOD}}(f_{dg}^{\mathcal{E}}), \tag{7}$$

where $f_{dg}^{\mathcal{E}}$ and $f_X^{P_{ID}}$ are the optimal domain-general and domain-specific predictions (Definitions 3-4).

The proof of Theorem 1 is provided in Appendix A.2. Additionally, our results hold for bounded features.

Theorem 1 suggests that a well-specified domain generalization split is one such that the OOD spurious correlation is misaligned with the ID spurious correlation and the variance spurious features are sufficiently controlled not to undo the effect of misalignment. *Additionally, for Gaussian Z_{spu}, these conditions are necessary, sufficient, and hold with certainty (Appendix A.3 Corollary 3).* Figure 1 demonstrates the accuracy of these conditions with simulated experiments.

Often, for evaluation, we are given a set of distributions, and we perform leave-one-domain-out ID and OOD splits. In the context of Definition 5, P_{ID} represents the mixture of ID distributions, and the test split is P_{OOD} (which can also be a mixture). Importantly, a Gaussian is sub-Gaussian and a finite mixture of sub-Gaussians is also sub-Gaussian (Appendix A.8 Lemma 3), so our results directly apply.

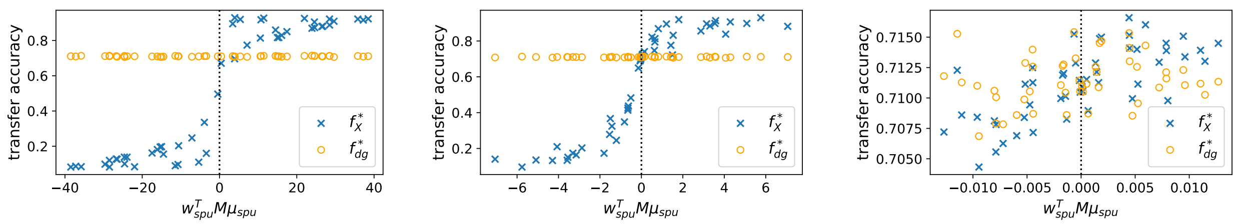
So far, we have shown that the learned domain-general and domain-specific predictors on a given training distribution differ, and the domain-general model only achieves higher OOD accuracy when spurious correlations ID and OOD are sufficiently misaligned. Next, we show that when we evaluate a set of diverse predictors on the ID and OOD test data and observe a weak or negative correlation between the predictors' ID and OOD accuracy, i.e., no *accuracy on the line*, this misalignment exists.

2.1 Accuracy on the Line

First, we define *accuracy on the line*, the correlation strength between in- and out-of-distribution accuracy.

Definition 6 (Accuracy on the Line (Miller et al., 2021))**.** *Define $a \in \mathbb{R}$, $\epsilon \geq 0$, and Φ^{-1} as the inverse Gaussian cumulative density function. The* correlation property *is defined as*

$$\left|\Phi^{-1}\left(acc_{P_{ID}}(f)\right) - a\cdot\Phi^{-1}\left(acc_{P_{OOD}}(f)\right)\right| \leq \epsilon\,\forall\, f. \tag{8}$$



(a) **All random variables are Gaussian.** When Theorem 1 conditions are satisfied OOD (x-axis: $\mathrm{w}_{\mathrm{spu}}^\top M\mu_{\mathrm{spu}} < 0$), the f_{dg}'s outperform f_{X}'s OOD. This result verifies that there needs to be sufficient misalignment between in- and out-of-distribution spurious correlations for the domain-general features to outperform the domain-specific models in OOD accuracy. Details on the experiments can be found in Appendix B.

(b) **$\mathbf{Z}_{\mathrm{spu}}^{\mathrm{ID}}$ is defined to be a mixture of 4 Gaussian distributions (sub-Gaussian) such that the test distributions are not a mixture of these 4 Gaussians.** The same conclusions in Figure 1a hold for sub-Gaussian random variables when the test distribution is an extrapolation.

(c) Rosenfeld et al. (2022b) studies domain interpolation, where the target domain is a mixture of the training domain; a variant of ERM is provably worst-shift optimal. **$\mathbf{Z}_{\mathrm{spu}}^{\mathrm{ID}}$ is defined to be a mixture of 4 Gaussian distributions such that the test distributions are a different mixture of the 4 Gaussians.** Overall, there is minimal difference in OOD accuracy between f_{dg} and f_{X} in this setting (domain interpolation).

Figure 1: f_{dg} are trained on (Z_{dg}, Y) pairs and f_{X} are trained on (X, Y) pairs from the same distribution. $X = Z_{\mathrm{dg}} \oplus Z_{\mathrm{spu}}^{\mathrm{ID}}$, where $\oplus$ is concatenation. $Z_{\mathrm{spu}}^{\mathrm{ID}}$ is Sub-Gaussian. We evaluate these models on 50 test distributions generated with randomly sampled M such that all other distributions are the same ID and OOD except test distribution $Z_{\mathrm{spu}}^{\mathrm{OOD}} = MZ_{\mathrm{spu}}^{\mathrm{ID}}$. Details on the experiments can be found in Appendix B.

If there exists an a such that $\epsilon \approx 0$, then there is a strong correlation between ID and OOD accuracy. If $a > 0$, then the correlation is positive, and if $a < 0$, the correlation is negative. We will call the setting where $a > 0$ positive accuracy on the line and $a < 0$ accuracy on the inverse line. We derive conditions for accuracy on the line to occur next.

Theorem 2 (Conditions for Accuracy on the Line)**.** *For brevity, consider Theorem 1's setting. Assume that there exists two constants $\epsilon_1, \epsilon_2 > 0$ such that*

$$\|M\mu_{spu} - \mu_{spu}\| \leq \epsilon_1 \quad \text{and} \quad \left\|w_{spu}^\top M\Sigma_{spu}M^\top w_{spu} - w_{spu}^\top \Sigma_{spu} w_{spu}\right\| \leq \epsilon_2. \tag{9}$$

Additionally, assume there exists a constant $B > 0$ such that for sufficiently small t, $\Pr(|f_X(X)| \leq t) \leq Bt$, and some $\alpha > 0$ such that $acc_P(f_X) \in [\alpha, 1-\alpha]$ and $acc_{P_M}(f_X) \in [\alpha, 1-\alpha]$. Φ is L-Lipschitz in $[\alpha, 1-\alpha]$.

Then for any classifier $f_X \in \mathcal{F}$, $\delta \in (0,1)$, and $a \in \mathbb{R}$, with probability at least $1-\delta$,

$$|\Phi^{-1}(acc_P(f_X)) - a\Phi^{-1}(acc_{P_M}(f_X))| \leq LB\left(\|w_{spu}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2}\right) + |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|, \tag{10}$$

The proof of Theorem 2 is provided in Appendix A.4.

Theorem 2 specifies conditions for there to be a strong correlation between ID and OOD accuracy. The strength of the correlation depends strongly on small shifts in the spurious correlation from train to test. For large deviations, e.g., when necessary for Theorem 1, the RHS of Equation 10 increases with ϵ_1, ϵ_2, and there will not be a strong correlation between ID and OOD accuracy. Corrolary 1 directly characterizes the tradeoff between accuracy on the line and the conditions for well-specified benchmarks in Theorem 1.

Corollary 1 (Tradeoff Between Accuracy on The Line and Well-Specification)**.** *Let $Z_{spu} \in \mathbb{R}^l$ be sub-Gaussian with parameter κ, mean μ_{spu}, and covariance Σ_{spu}. Fix $w_{spu} \in \mathbb{R}^l$ such that $w_{spu}^\top \mu_{spu} > 0$. Suppose $M \in \mathbb{R}^{l\times l}$ satisfies the* spurious correlation reversal condition *for some margin $\gamma > 0$, i.e.,*

$$w_{spu}^\top(M\mu_{spu}) + \sqrt{2\kappa^2 w_{spu}^\top M\Sigma_{spu}M^\top w_{spu}\log(1/\delta)} \leq -\gamma < 0. \tag{11}$$

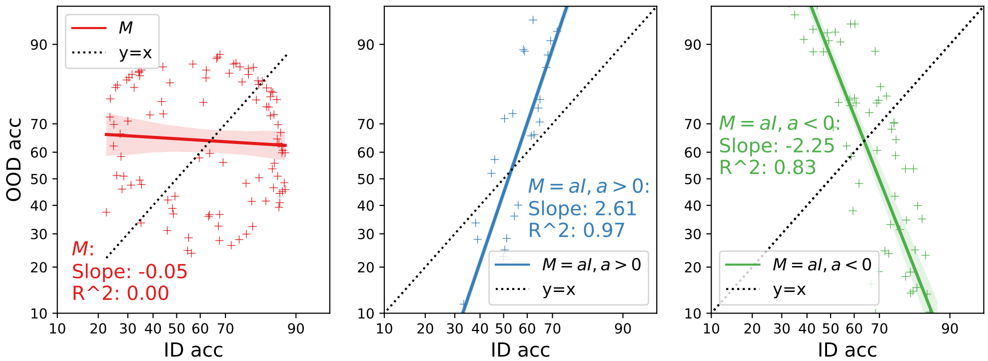


Figure 2: ID vs. OOD accuracy on probit scale. When M satisfies Theorem 1's conditions, the accuracy on the line phenomenon does not occur. For $M_{\text{ID}} = I$ and $M_{\text{OOD}} = aI$, where a is allowed to vary, we observe accuracy on the line. When $a < 0$, we have the spurious correlation reversal condition and have accuracy on the inverse line, where these is a strong but negative correlation between in- and out-of-distribution accuracy. Notably, ID/OOD splits with accuracy on the inverse line are well-specified. More experimental details can be found in Appendix B.

Assume also that there exists some $\alpha > 0$ such that $acc_P(f_X)$, $a \cdot acc_{P_M}(f_X) \in [\alpha, 1-\alpha]$. Then with probability at least $1-\delta$,

$$|\Phi^{-1}(acc_P(f_X)) - a \cdot \Phi^{-1}(acc_{P_M}(f_X))| \geq C\kappa \|w_{spu}\| \sqrt{\log(1/\delta)} \|M\mu_{spu} - \mu_{spu}\| + \zeta, \tag{12}$$

for some constant $C > 0$ (depending on α and the local slope of Φ^{-1}) where $a \in \mathbb{R}$, $\zeta = |1 - a| \max_{x \in [\alpha, 1-\alpha]} |\Phi^{-1}(x)|$.

Corollary 1 shows that, generally, one would not expect the conditions for well-specified splits to hold simultaneously while maintaining accuracy on the line.

Corollary 2 (Informal)**.** *In a measure–theoretic sense with respect to Lebesgue measure, the set of shifts with $\epsilon \to 0$ (Definition 6), i.e., perfect accuracy on the line, is misspecified almost everywhere. Furthermore, the Lebesgue measure of shifts that are well-specified grows monotonically with ϵ (i.e., with weaker accuracy on the line). Full result and proof are provided in Appendix A.6.*

In Appendix A.7, we construct an examples of the shifts implied in Corollary 2, where both the conditions for well-specified splits and accuracy on the line hold.

The trade-off established by Theorem 1 is fundamental: achieving a negative margin can require M to shift μ_{spu} substantially, as measured by $\|M\mu_{\text{spu}} - \mu_{\text{spu}}\|$. The magnitude of such M is large when there is a strong in-distribution correlation w_{spu}; such a strong correlation can lead to the most catastrophic failures under worst-case shifts. This necessarily impacts the accuracy difference through the lower bound of Theorem 1, demonstrating the fact that sufficient spurious correlation reversal results in decreased OOD performance when the in-distribution models rely on spurious correlations. This decrease implies large ϵ in Definition 6 when $a > 0$. Hence, with high probability, ID/OOD splits that are well-specified do not exhibit a strong positive correlation between in- and out-of-distribution accuracy for arbitrary classifiers. Then, one would expect a weak correlation or a strong correlation with $a < 0$ (*accuracy on the inverse line*). *Accuracy on the line gives a test to eliminate current misspecified benchmarks as well as identify future misspecified benchmarks.*

Theorems 1 and Corollary 1-2 establish the link between *accuracy on the line* and the utility of domain generalization benchmarks. With high probability, accuracy on the line generally does not occur when conditions for well-specified domain generalization benchmarks (Theorem 1) are satisfied—Figure 2.

Limitations of Accuracy on the Line Benchmarks. Our results suggest that datasets with accuracy on the line require more thought and attention to ensure that they are well-specified domain generalization benchmarks. In contrast, benchmarks with *accuracy on the inverse line*, or weak correlation between ID and OOD accuracy, are better-suited to benchmark domain generalization.

3 Related Work

Two influential domain generalization benchmark suites are DomainBed (Gulrajani & Lopez-Paz, 2020a) and WILDS (Koh et al., 2021). **DomainBed** is a collection of object recognition domain generalization benchmarks. For example, PACS (Khosla et al., 2012; Li et al., 2017) includes images of seven classes across four domains: Photos, Art Paintings, Cartoons, and Sketches. Another benchmark is ColoredMNIST (Arjovsky et al., 2019), a semi-synthetic binary classification variation of MNIST (Deng, 2012), which introduces color as a spurious correlation and defines domains by specific color-label associations. Gulrajani & Lopez-Paz (2020a) found that empirical risk minimization achieved the best transfer performance compared to state-of-the-art domain generalization algorithms across PACS, ColoredMNIST, and other DomainBed benchmarks. They also observed differences in their findings based on their choice of model selection.

WILDS was designed to better represent real-world shifts across vision and language. For example, Camelyon17 (Zech et al., 2018; AlBadawy et al., 2018) includes images of tissue that may contain tumor tissue (classes) from different hospitals with varying conditions (domains). CivilComments Borkan et al. (2019) includes comments to an online article that may be toxic (class) for different subpopulations defined by demographic identities (domains). These benchmarks illustrate the suite's focus on practical and natural real-world applications. However, even across WILDS benchmarks, no state-of-the-art methods have demonstrated consistent superiority over ERM (Koh et al., 2021).

For **Subpopulation shift**, which we consider a special case of the broader domain generalization task, benchmarks are designed specifically to evaluate distribution shift robustness in scenarios where spurious correlations lead to worse performance on underrepresented subgroups out-of-distribution. For example, the Waterbirds benchmark (Sagawa et al., 2019) introduces a spurious correlation between bird species and backgrounds, such as waterbirds predominantly appearing in water environments. Models that rely on backgrounds rather than bird features when predicting bird type (classes) generalize poorly to new domains where backgrounds are urban areas—birds in urban backgrounds are undersampled in the training data.

Benchmarks addressing **other types of distribution shifts** where domain generalization is desired have also been proposed, e.g., temporal shifts (Yao et al., 2022; Joshi et al., 2023; Zhang et al., 2023).

3.1 Accuracy on the Line

The accuracy on the line phenomena has been observed empirically in previous work for many benchmarks in these suites (Recht et al., 2019; Miller et al., 2021; 2020; Taori et al., 2020; Baek et al., 2022; Saxena et al., 2024). However, Liu et al. (2023a) identify real-world tabular datasets with weak or negative linear correlation, and Teney et al. (2023) identify non-tabular real-world datasets where ID and OOD performance exhibit other patterns between in- and out-of-distribution accuracy beyond strongly positive and linear. Notably, they also found that correlations can be inverted in some datasets. Sanyal et al. (2024) derive noise conditions to achieve shifts with accuracy on the inverse line. Our work uniquely specifies the implications of accuracy on the line (or lack thereof) on the utility of datasets as domain generalization benchmarks.

Other works have studied conditions where the empirical risk minimizer for observed distributions is sufficient for domain generalization. Rosenfeld et al. (2022b) show this to be the case under domain interpolation, i.e., OOD distributions are convex combinations of observed distributions (in an online setting). Additionally, when domain-general features are fully informative, i.e., spurious correlations are redundant, and under some conditions, Ahuja et al. (2021) also show this to be the case. Our results corroborate their findings.

This notion of worst-case stress testing to establish robustness under distribution shift is not new. When evaluating the out-of-distribution generalization of deep learning methods developed on the ImageNet task (Deng et al., 2009) (*ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* (Russakovsky et al., 2015)), Kornblith et al. (2019) assess generalization by applying said methods to contemporary datasets with presumably different distributions than ImageNet, such as CIFAR-10 (Krizhevsky, 2009). However, the datasets they investigated could be considered quite similar in distribution to ImageNet. Alternatively, Salaudeen & Hardt (2024) adversarially constructed a dataset, ImageNot, designed to shift spurious correlations present in the original ImageNet construction strongly. Notably, the findings of Salaudeen & Koyejo (2022) align with those of Kornblith et al. (2019) despite the deliberately adversarial construction of ImageNot. We argue that constructing datasets with such adversarial spurious correlation shifts is essential for rigorously probing a model's use of spurious correlations.

More generally, other works have proposed alternative approaches to developing domain generalization benchmarks. Satisfying our conditions introduces an additional necessary dimension for creating more meaningful evaluations. For example, Zhang et al. (2023) argues that many existing benchmarks are limited by having too few domains and overly simplistic settings, which restrict their ability to simulate the significant distribution shifts observed in real-world scenarios. Similarly, Lynch et al. (2023) contend that benchmarks inadequately capture the complex, many-to-many spurious correlations that can arise in practical applications.

Like previous work, we next investigate the accuracy of the line properties of state-of-the-art domain generalization benchmarks to take an inventory of well-specified datasets.

4 Empirical Results

We evaluate the correlation between in-domain (ID) and out-of-domain (OOD) accuracy for benchmarks in the popular DomainBed (Gulrajani & Lopez-Paz, 2020a) and WILDS (Koh et al., 2021) benchmark suites, as well as subpopulation shift benchmarks, e.g., WaterBirds (Sagawa et al., 2019).

Datasets. Specifically, our results include the following datasets: **Camelyon** (Bandi et al., 2018; Koh et al., 2021), **CivilComments** (Borkan et al., 2019; Koh et al., 2021), **ColoredMNIST** (Arjovsky et al., 2019; Gulrajani & Lopez-Paz, 2020a), **Covid-CXR** (Alzate-Grisales et al., 2022; Cohen et al., 2020b; Tabik et al., 2020; Tahir et al., 2021; Suwalska et al., 2023), **FMoW** (Christie et al., 2018; Koh et al., 2021), **PACS** (Li et al., 2017; Gulrajani & Lopez-Paz, 2020a), **Spawrious** (Lynch et al., 2023), **TerraIncognita** (Beery et al., 2018; Gulrajani & Lopez-Paz, 2020a), and **Waterbirds** (Sagawa et al., 2019).

Model Architectures. For vision datasets, we leverage pretrained deep learning architectures such as **ResNet-18/50** (He et al., 2016), **DenseNet-121** (Huang et al., 2017), **Vision Transformers** (Dosovitskiy et al., 2020), and **ConvNeXt-Tiny** (Liu et al., 2022). For language datasets, we utilize pretrained embeddings from **BERT** Jacob et al. (2019) and **DistilBERT** Sanh et al. (2020), and apply lower-capacity machine learning models, such as logistic regression, for downstream classification tasks.

Experimental Setup. The benchmarks we consider include a set of domains—distinct data distributions. As standard in the literature (Gulrajani & Lopez-Paz, 2020a; Koh et al., 2021), the in-distribution (ID) data is defined as a mixture of a subset of the data domains, and the out-of-distribution (OOD) data is not included in the training domains, i.e., we perform a leave-one-domain-out ID/OOD splits for our experiments, where we train on all but one domain and use the left-out domain as OOD. We generate models for our experiments by varying model and training hyperparameters for each architecture, including the number of training epochs (Appendix C Table 2). Our experiments include training these models end-to-end with varying hyperparameters and data augmentations, as well as pretraining and transfer learning.

Table 1 highlights the prevalence of widely-used domain generalization benchmarks with accuracy on the line, a signature of potential misspecification, while Figure 3 qualitatively illustrates benchmarks with weak or strongly negative correlation between in and out-of-distribution accuracy. We provide a detailed account of our experiments, along with benchmark-specific discussions, in Appendix C.

4.1 Findings

We find that, primarily, semisynthetic datasets satisfy our derived conditions. Particularly, semisynthetic here means real-world datasets that either (i) have artificial spurious correlations introduced (ColoredMNIST, Spawrious, and Waterbirds) and (ii) have some selection process that introduces spurious correlations (CivilComments). The CXR datasets exhibit a relationship that Teney et al. (2024) refers to as 'Not Transfer,' where the OOD accuracy is near constant despite high variance in the ID accuracy. Our results further echo

Table 1: We train on a set of ID distributions and test on a left-out OOD distribution. We present the Pearson R of ID and OOD probit-transformed accuracies and the slope and intercept of OOD accuracy regressed on ID accuracy. (*) OOD for waterbirds refers to the group where $y = 0$ and $a = 0$. The ID dataset is the mixture of groups at train time. Additional datasets and analysis are provided in Appendix C, which also includes complete tables with all splits for each dataset. With a threshold of $|R| < 0.5$, only a subset of datasets satisfy our derived conditions.

Dataset	**OOD**	$\|R\| < 0.5$	**slope**	**offset**	**R**	**p-value**	**std error**
ColoredMNIST	Env 2 acc	✓	-1.56	0.47	-0.74	0.00	0.01
CXR	Env 1 acc	✓	-0.60	0.56	-0.48	0.00	0.03
SpawriousO2O hard	Env 0 acc	✓	0.32	-0.21	0.50	0.00	0.05
SpawriousM2M hard	Env 0 acc	✓	0.76	-0.26	0.94	0.00	0.01
SpawriousO2O easy	Env 0 acc	✗	0.48	-0.29	0.74	0.00	0.04
SpawriousM2M easy	Env 0 acc	✗	0.34	0.26	0.60	0.00	0.00
PACS	Env 1 acc	✗	0.68	-0.68	0.84	0.00	0.01
TerraIncognita	Env 1 acc	✗	0.83	-1.41	0.74	0.00	0.02
Camelyon	Env 2 acc	✗	0.62	0.49	0.78	0.00	0.01
FMoW	Env 5 acc	✗	0.76	-0.61	0.87	0.00	0.01
CivilComments	Env 1 acc	✓	-0.49	0.16	-0.47	0.00	0.03
WaterBirds	Env 0 (*) acc	✓	-0.13	1.58	-0.13	0.00	0.03

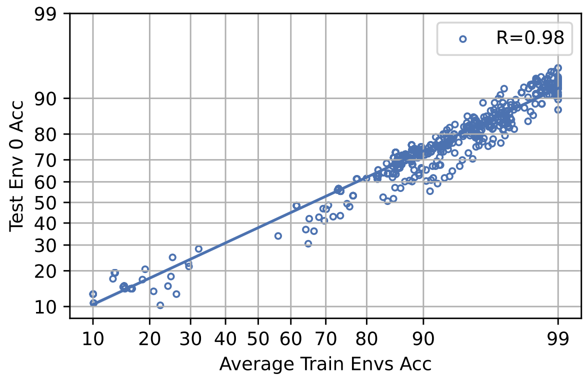
Teney et al. (2024)'s notes on potentially misleading advice from past studies—particularly the focus on improving ID performance to improve OOD robustness (Wenzel et al., 2022), which our work suggests is only a reasonable strategy in settings where distribution shifts are relatively simple and weak.

Our results and findings should not be surprising, given that spurious correlations we aim to mitigate for decision-making tend not to be localized (Teresa-Morales et al., 2022). For instance, correlations between gender and occupation likely persist across naturally collected datasets. Hence, they may not harm performance across different data sources. We find that the subpopulation shift benchmarks we assess often have the desired properties derived in this work. Notably, these benchmarks are (i) constructed such the spurious correlation from training is no longer accurate at testing and (ii) evaluated based on worst-group (worst-case) performance. Our work shows that in less concisely defined domain generalization contexts, such types of shifts are also necessary. We provide detailed results and discussion in Appendix C.

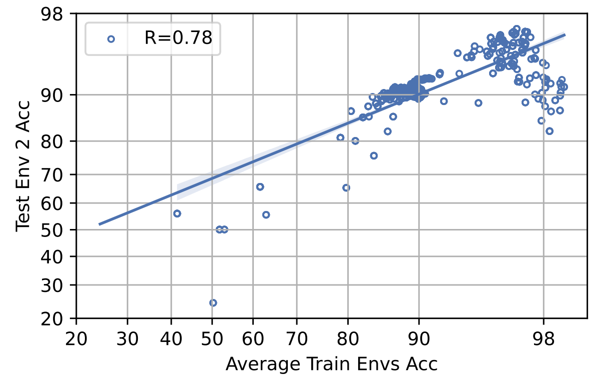
4.2 Discussion

Domain generalization considers worst-case shifts (Arjovsky et al., 2019; Rosenfeld et al., 2022b). While worst-case shifts may have a trade-off with average utility (Salaudeen & Koyejo, 2024; Miller, 2024), it is crucial in fairness-sensitive and high-stakes applications like healthcare (Angwin et al., 2016; Chen et al., 2019; Oakden-Rayner et al., 2020). However, domain generalization benchmarks have not been rigorously assessed for their utility in such scenarios. We address this gap by deriving conditions for well-specified benchmarks, helping practitioners align benchmark choices with their goals. We showed that well-specified benchmarks, i.e., are aligned with worst-case shifts, will not have accuracy on the line. Hence, benchmark users and curators with a goal of addressing such shifts should seek out benchmarks with weak or negative ID and OOD accuracy correlation.

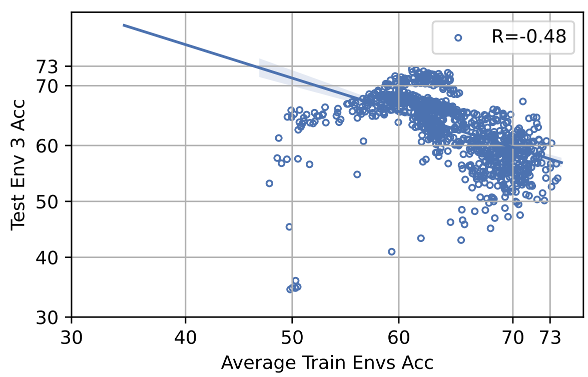
Many benchmark curators define their intended scope carefully. For example, WILDS (Koh et al., 2021) focuses on real-world rather than worst-case shifts. Our findings often validate these intended scopes, yet benchmark users may not always adhere to them. To clarify benchmark suitability, we categorize benchmarks into (i) worst-case and (ii) natural shifts (Koh et al., 2021; Taori et al., 2020), emphasizing that worst-case benchmarks are particularly valuable for auditing uninterpretable predictors, such as detecting demographic biases in model decisions (Ferrer et al., 2021), whereas natural shifts may suffice when prioritizing average OOD performance. Our conditions enable assessing whether spurious correlations (e.g., race in Chest X-Rays (Gichoya et al., 2022)) impact predictions. For robust algorithm development, both worst-case and



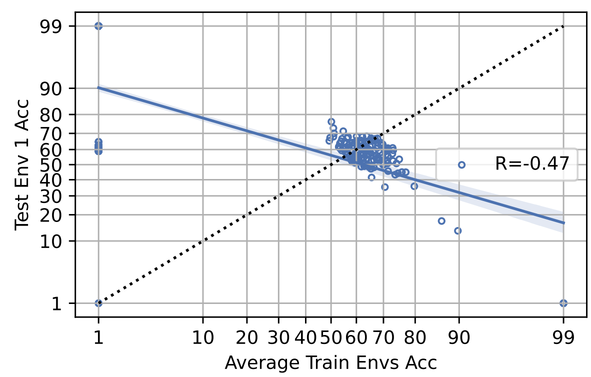
(a) PACS (Env 0)



(b) Camelyon (Env 2)



(c) Covid-CXR (Env 3)



(d) CivilComments (Env 1).

Figure 3: We show some ID/OOD splits of popular domain-generalization benchmarks with a strong positive, weak, or strong negative correlation between in-distribution and out-of-distribution accuracy. Our results suggest that algorithms that consistently provide models with the best transfer accuracies for these splits are at least partially successful in removing spurious correlations. For Camelyon (b), within some accuracy range, we have accuracy on the inverse line, indicating the importance of a qualitative assessment of these trends.

natural shifts should be considered to ensure broad applicability. Additionally, new spurious correlation benchmarks should undergo *accuracy on the line* evaluations to support their reliability—ideally, there is no positive accuracy on the line. Finally, domain expertise is critical in defining spurious correlations: a narrow potential distribution set may limit robustness, while an overly broad definition of potential distributions can unnecessarily reduce utility (Shen et al., 2024). For instance, Chiou et al. (2024) report that training a model on multiple recording sessions degrades OOD (new session) brain-computer interface (BCI) classification performance compared to training on single-session data.

Constructing Benchmarks Without Positive Accuracy on the Line. Indeed, some of the datasets we study empirically that satisfy our desired conditions are primarily semisynthetic (e.g., ColoredMNIST, Spawrious, Waterbirds). In contrast, datasets such as CivilComments, Covid-CXR, and WILDSCamelyon, which have in-distribution (ID) and out-of-distribution (OOD) splits, do not exhibit positive accuracy on the line. This suggests that careful dataset collection is likely necessary to obtain naturally occurring datasets without positive accuracy on the line.

One approach to constructing such datasets is to identify settings where a natural experiment or intervention has occurred. For example, the Covid-CXR dataset leverages the natural intervention of the pandemic and regional variations. Without such interventions, there may be little reason to expect that spurious correlations would fail to generalize in the datasets that are most readily available.

Qualitatively Assessing Accuracy on the Line. For the Camelyon dataset (Section C.6), we observe a shift in correlation patterns based on model accuracy. Specifically, models with greater than 90% accuracy exhibit a negative correlation between ID and OOD performance, whereas models below this threshold show a positive correlation. Since accuracy on the line (Definition 2) is a global property, this dataset does not meet the criteria for accuracy on the line since there is a strong deviation from the positive correlation in some regimes. This underscores a key limitation: evaluating overall correlation may not sufficiently identify well-specified benchmarks. While this approach is robust against false negatives of misspecification, it may introduce false positives, necessitating qualitative inspection as a complementary assessment tool.

On the Slope vs. the Correlation Coefficient. In this work, we focus on the correlation between ID and OOD accuracy rather than the slope between the two quantities. Recall that for a set of pairs x, y, which have a Pearson R correlation of r, the slope when y is regressed on x is $s = r * (\sigma_y/\sigma_x)$. s depends on variances σ_y and σ_x, which may or may not be relevant to the spurious correlation problem. The slope depends strongly on the definition of $\mathcal{F}$ (Appendix C). This observation is also related to limitations in attributing accuracy drop across distribution only to distributional systematic bias (Salaudeen & Hardt, 2024). While s is informative, its relationship with our derived conditions is not as obvious.

Reconciling Worst-Case and Average-Case Generalization. A strict worst-case approach to generalization can conflict with developing locally (sites) beneficial models, particularly in healthcare (Futoma et al., 2020; Miller, 2024). However, failures even within the same site suggest this tension is unavoidable (Oakden-Rayner et al., 2020), as spurious correlations remain brittle due to other non-local factors like temporal drifts (Ji et al., 2023) or interventions (Birkmeyer et al., 2020). Reliable worst-case robustness benchmarks remain essential, even with a narrower scope. Still, when worst-case focus severely impacts average utility, additional evaluation on alternative benchmarks is warranted. A practical approach is narrowing the model's deployment scope to a smaller set of distributions, which can improve worst-case performance (e.g., when the shifts in the smaller set are weaker) without sacrificing as much average utility. However, practically, maintaining reliable predictions may require scope-dependent model monitoring and updates.

Implications on Key Domain Generalization Benchmarking Practices. ID/OOD splits within the same dataset can exhibit varying Pearson R correlations, meaning some splits provide more reliable benchmarks than others. However, **averaging** over all ID/OOD splits (Gulrajani & Lopez-Paz, 2020a) dilutes this reliability, particularly when only one split is well-specified. The issue worsens when averaging across multiple datasets to compare domain generalization methods (Gulrajani & Lopez-Paz, 2020a). In subpopulation shifts, **worst-group accuracy** is the standard evaluation metric (Koh et al., 2021); adopting a similar norm more generally for domain generalization improves the robustness of evaluation.

A related issue arises in **model selection** via cross-validation, where selecting models based on held-out accuracy—whether IID or a held-out domain—can lead to overfitting to spurious correlations specific to that set. Alternative selection criteria are implied conditional independencies (Salaudeen & Koyejo, 2024), cross-risk minimization (Pezeshki et al., 2023), and confidence-based ensemble aggregation (Chen et al., 2023). However, model selection under distribution shifts remains a challenge.

Implications on Benchmarking Causal Representation Learning which aims to uncover underlying causal structures. One approach to this is *independent causal mechanisms* (Pearl, 2009; Schölkopf et al., 2021), which has motivated many domain generalization algorithms (Arjovsky et al., 2019; Peters et al., 2016). Since both tasks require distinguishing stable from spurious correlations, our results on evaluating domain generalization also apply to benchmarking causal representation learning. Specifically, when assessing models—including disentangled causal models—based on OOD accuracy, our framework helps determine when domain generalization reliably reflects success in learning causal representations (Salaudeen et al., 2024)—we discuss this further in Appendix D.

Implications on Benchmarking Algorithmic fairness which aims to mitigate biases that cause disparate performance across demographic groups; some definitions of fairness are closely linked to domain generalization (Creager et al., 2021). Group sufficiency particularly aligns with the principle of invariance (Chouldechova, 2017; Liu et al., 2019). We emphasize a straightforward but key insight: when using OOD accuracy to benchmark whether models avoid relying on group information, the benchmark must ensure that group information hinders out-of-distribution performance. In this case, a strong positive correlation between training and worst-group test accuracy suggests that group information generalizes. In contrast, a weak or negative correlation between ID and OOD accuracy is preferable.

Modern **foundation models** are susceptible to spurious correlations (Alabdulmohsin et al., 2024; Zhu et al., 2023; Gerych et al., 2024; Hamidieh et al., 2024). Analyzing the Civil Comments dataset, we find that spurious correlation shifts in language datasets exhibit similar patterns to vision datasets within our framework, showing strong positive, negative, and weak correlations between ID and OOD accuracy. Furthermore, Saxena et al. (2024) report strong positive correlations in large language models for predictive tasks (Q/A) under distribution shift. Our results highlight that the benchmark conditions we establish are also crucial for evaluating spurious correlations in foundation models.

4.3 Limitations and Future Work

While we have developed probabilistic sufficient conditions for well-specified benchmarks, we derive necessary and sufficient conditions only under the Gaussianity assumption of features. Future work includes exploring whether these conditions are necessary for non-Gaussian features. Demonstrating this may require developing analytical techniques distinct from those applied in this work. Additionally, Yang et al. (2023) demonstrates that while accuracy on the line may hold, other metrics or a combination of metrics may not simultaneously have the same strong positive linear trend. Investigating metrics other than accuracy is left for future work.

Another direction for future work is to develop a more robust automated method for assessing accuracy on the line, beyond simply computing correlation across all training accuracy levels. Currently, qualitative assessment remains necessary to account for cases where models with higher training accuracy exhibit negative accuracy on the line. Empirically, we found that standard change-point detection methods (Killick et al., 2012) are highly sensitive to noise in accuracy measurements. However, incorporating more robust heuristics could improve their reliability, making them a viable approach for automation.

Additionally, while we have evaluated a wide variety of models in this work, continuing to collect data points of ID/OOD accuracies for benchmarks improves the accuracy of the true relationship. Finally, curating additional benchmarks without accuracy on the line, drawn from diverse, real-world scenarios with high-dimensional spurious features, is left for future work. This work, along with others (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021), has characterized this property for a variety of popular benchmarks.

5 Conclusion

Robustness to spurious correlations under worst-case distribution shifts is a critical challenge in machine learning, essential for ensuring the reliability and fairness of models. In this work, we identify significant limitations in current benchmarks designed to address this problem. Specifically, many state-of-the-art benchmarks, which evaluate OOD accuracy by training models on an in-distribution (ID) split and testing on an out-of-distribution (OOD) split, fail to guarantee that models free of spurious correlations will transfer better. We define a benchmark as *well-specified* if such a guarantee exists.

Previous work observed that many benchmarks exhibit the phenomenon of *accuracy on the line*, where improved ID performance directly correlates with improved out-of-distribution performance. Our theoretical findings suggest that this behavior indicates that such benchmarks are misspecified for evaluating domain generalization and emphasize the importance of prioritizing benchmarks that do not exhibit accuracy on the line when addressing worst-case distribution shifts. We aim to provide a clearer path toward developing models robust to spurious correlations by addressing the evaluation ambiguity.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*, 2024.

Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

Jesús Alejandro Alzate-Grisales, Alejandro Mora-Rubio, Harold Brayan Arteaga-Arteaga, Mario Alejandro Bravo-Ortiz, Daniel Arias-Garzón, Luis Humberto López-Murillo, Esteban Mercado-Ruiz, Juan Pablo Villa-Pulgarin, Oscar Cardona-Morales, Simon Orozco-Arias, et al. Cov-caldas: A new covid-19 chest x-ray dataset from state of caldas-colombia. *Scientific Data*, 9(1):757, 2022.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

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

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly Stachenfeld, Pushmeet Kohli, Peter Battaglia, and Jessica Hamrick. Structured agents for physical construction. In *International conference on machine learning*, pp. 464–474. PMLR, 2019.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

John D Birkmeyer, Amber Barnato, Nancy Birkmeyer, Robert Bessler, and Jonathan Skinner. The impact of the covid-19 pandemic on hospital admissions in the united states: study examines trends in us hospital admissions during the covid-19 pandemic. *Health Affairs*, 39(11):2010–2017, 2020.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Annie S Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. Confidence-based model selection: When to take shortcuts for subpopulation shifts. *arXiv preprint arXiv:2306.11120*, 2023.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

Nicole Chiou, Mehmet Günal, Sanmi Koyejo, David Perpetuini, Antonio Maria Chiarelli, Kathy A. Low, Monica Fabiani, and Gabriele Gratton. Single-trial detection and classification of event-related optical signals for a brain–computer interface application. *Bioengineering 2024, Vol. 11, Page 781*, 11:781, 8 2024. ISSN 2306-5354. doi: 10.3390/BIOENGINEERING11080781. URL `https://www.mdpi.com/2306-5354/11/8/781/htmhttps://www.mdpi.com/2306-5354/11/8/781`.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. 2020a. URL `https://arxiv.org/abs/2002.02497`.

Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020b.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

Xavier Ferrer, Tom Van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2):72–80, 2021.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9): e489–e492, 2020.

Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad-Javad Darvishi-Bayazi, Pooneh Mousavi, Guillaume Dumas, and Irina Rish. Woods: Benchmarks for out-of-distribution generalization in time series. *arXiv preprint arXiv:2203.09978*, 2022.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 36:53385–53432, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. *arXiv preprint arXiv:2411.04420*, 2024.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020a.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. 2020b. URL `https://arxiv.org/abs/2007.01434`.

Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pp. iii–115, 1944.

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 547–561, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Kevin D Hoover. The logic of causal inference: Econometrics and the conditional analysis of causation. *Economics & Philosophy*, 6(2):207–234, 1990.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019. URL `https://arxiv.org/abs/1810.04805`.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Christina X Ji, Ahmed M Alaa, and David Sontag. Large-scale study of temporal shift in health insurance claims. In *Conference on Health, Inference, and Learning*, pp. 243–278. PMLR, 2023.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. *arXiv preprint arXiv:2306.11957*, 2023.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 158–171. Springer, 2012.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL `https://api.semanticscholar.org/CorpusID:18268744`.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. icitris: Causal representation learning for instantaneous temporal effects. In *UAI 2022 Workshop on Causal Representation Learning*, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pp. 13557–13603. PMLR, 2022b.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51371–51408. Curran Associates, Inc., 2023a. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/a134eaebd55b7406ff29cd75d5f1a622-Paper-Datasets_and_Benchmarks.pdf`.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pp. 4051–4060. PMLR, 2019.

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Conference on Causal Learning and Reasoning*, pp. 553–573. PMLR, 2023b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pp. 662–691. PMLR, 2023.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6979–6987, 2017.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8046–8056, 2022.

Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

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

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 739–766. PMLR, 5 2022. URL `http://arxiv.org/abs/2105.06422`.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International conference on machine learning*, pp. 6905–6916. PMLR, 2020.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Katharine Miller. Healthcare algorithms don't always need to be generalizable. *Stanford HAI*, 2024. URL `https://hai.stanford.edu/news/healthcare-algorithms-dont-always-need-be-generalizable`. Accessed: 2024-07-01.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Vivian Yvonne Nastl and Moritz Hardt. Do causal predictors generalize better to new domains? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

J Pearl. *Causality.* Cambridge university press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with xrm. *arXiv preprint arXiv:2309.16748*, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022a.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2641–2657. PMLR, 2022b.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Olawale Salaudeen and Moritz Hardt. Imagenot: A contrast with imagenet preserves model rankings. *arXiv preprint arXiv:2404.02112*, 2024.

Olawale Salaudeen and Sanmi Koyejo. Causally inspired regularization enables domain general representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3124–3132. PMLR, 2024.

Olawale Elijah Salaudeen and Oluwasanmi O Koyejo. Exploiting causal chains for domain generalization. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

Olawale Elijah Salaudeen, Nicole Chiou, and Sanmi Koyejo. On domain generalization datasets as proxy benchmarks for causal representation learning. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pp. 8459–8468. PMLR, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 2020. URL `https://arxiv.org/abs/1910.01108`.

Amartya Sanyal, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf. Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation. *arXiv preprint arXiv:2406.19049*, 2024.

Rahul Saxena, Taeyoun Kim, Aman Mehra, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Predicting the performance of foundation models via agreement-on-the-line. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y. Chen. The data addition dilemma, 2024. URL `https://arxiv.org/abs/2408.04154`.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Aleksandra Suwalska, Joanna Tobiasz, Wojciech Prazuch, Marek Socha, Pawel Foszner, Damian Piotrowski, Katarzyna Gruszczynska, Magdalena Sliwinska, Jerzy Walecki, Tadeusz Popiela, Grzegorz Przybylski, Mateusz Nowak, Piotr Fiedor, Malgorzata Pawlowska, Robert Flisiak, Krzysztof Simon, Gabriela Zapolska, Barbara Gizycka, Edyta Szurowska, Agnieszka Oronowicz-Jaskowiak, Bogumil Golebiewski, Mateusz Rataj, Przemyslaw Chmielarz, Adrianna Tur, Grzegorz Drabik, Justyna Kozub, and et al. Kozanecka. Polcovid: a multicenter multiclass chest x-ray database (poland, 2020–2021). *Scientific Data*, 10(1):348, Jun 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02229-5. URL `https://doi.org/10.1038/s41597-023-02229-5`.

S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and F. Herrera. Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3595–3605, 2020. doi: https://ieeexplore.ieee.org/document/9254002.

Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.105002. URL `https://www.sciencedirect.com/science/article/pii/S0010482521007964`.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 71703–71722. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/e304d374c85e385eb217ed4a025b6b63-Paper-Conference.pdf`.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Cristina Teresa-Morales, Margarita Rodríguez-Pérez, Miriam Araujo-Hernández, and Carmen Feria-Ramírez. Current stereotypes associated with nursing and nursing professionals: An integrative review. *International journal of environmental research and public health*, 19(13):7640, 2022.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Jiaxuan Wang, Sarah Jabbour, Maggie Makar, Michael Sjoding, and Jenna Wiens. Learning concept credible models for mitigating shortcuts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33343–33356, 12 2022a.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022b.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324, 2022.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16036–16047, 2023.

Jiayun Zheng and Maggie Makar. Causally motivated multi-shortcut identification and removal. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12800–12812, 12 2022.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding, 2016. URL `https://arxiv.org/abs/1610.02055`.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. Debiased fine-tuning for vision-language models by prompt regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3834–3842, 2023.

Appendix Table of Contents

A Theoretical Results

A.1 Proof of Lemma 1—Domain-Specific Models have Lower In-Domain Error under Partially Informative Domain-General Features

Assume Partially Informative Domain-General Features (Assumption 2) and strongly convex ℓ.

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim P}\big[\ell(f(X),Y)\big] < \min_{f \in \mathcal{F}_{\text{dg}}} \mathbb{E}_{(X,Y)\sim P}\big[\ell(f(X),Y)\big], \tag{13}$$

where $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ where $f(x) = \sigma((\mathrm{w})^\top x) = \sigma(\mathrm{w}_{\text{dg}}^\top z_{\text{dg}} + \mathrm{w}_{\text{spu}}^\top z_{\text{spu}})$, $f \in \mathcal{F}$. For $f \in \mathcal{F}_{\text{dg}}$, $f(x) = \sigma((\mathrm{w})^\top x) = \sigma(\mathrm{w}_{\text{dg}}^\top z_{\text{dg}})$.

Proof. From the PIDGF assumption, a model that uses both domain-general and spurious features is more expressive than one that does not. By the Bayes optimality of w^*, any w achieving the same risk agrees with w^* almost everywhere, i.e.,

$$\mu\left(\left\{x \in \mathcal{X} \mid \sigma((\mathrm{w}^*)^\top x) \neq \sigma(\mathrm{w}^\top x)\right\}\right) = 0,$$

where $\sigma(\cdot)$ is the sigmoid function and μ denotes the Lebesgue measure on $\mathcal{X}$.

Let $x = z_{\mathrm{dg}} \oplus z_{\mathrm{spu}}$ and $\mathrm{w} = \mathrm{w}_{\mathrm{dg}} \oplus \mathrm{w}_{\mathrm{spu}}$ such that

$$\sigma((\mathrm{w}^*)^\top x) = \sigma(\mathrm{w}_{\mathrm{dg}}^\top z_{\mathrm{dg}} + \mathrm{w}_{\mathrm{spu}}^\top z_{\mathrm{spu}}).$$

If we only consider values of x where $\mathrm{w}_{\mathrm{dg}}^\top z_{\mathrm{dg}} \neq 0$, then without loss of generality we have that

$$\sigma((\mathrm{w}^*)^\top x) = \sigma(\mathrm{w}_{\mathrm{dg}}^\top z_{\mathrm{dg}} + \mathrm{w}_{\mathrm{spu}}^\top z_{\mathrm{spu}}) \neq \sigma(\mathrm{w}_{\mathrm{dg}}^\top z_{\mathrm{dg}}).$$

Given that

$$\mu\left(\left\{x \mid \mathrm{w}_{\mathrm{dg}}^\top z_{\mathrm{dg}} \neq 0\right\}\right) > 0,$$

the risk of $\sigma(\mathrm{w}_{\mathrm{dg}}^\top z_{\mathrm{dg}})$ is strictly greater than that of $\sigma((\mathrm{w}^*)^\top x)$. Equation 13 follows from the strong convexity of the loss. $\square$

A.2 Proof of Theorem 1—Conditions For Well-Specified Domain Generalization Benchmark Splits

Assume $Z_{\mathrm{spu}}^{\mathrm{ID}}$ is sub-Gaussian with mean μ_{spu}, variance Σ_{spu}, and parameter κ. Define $M \in \mathbb{R}^{l \times l}$ such that $Z_{\mathrm{spu}}^{\mathrm{OOD}} = M Z_{\mathrm{spu}}$ where in-distribution $Z_{\mathrm{spu}}^{\mathrm{OOD}} \sim P_{\mathrm{OOD}}$ and out-of-distribution $Z_{\mathrm{spu}}^{\mathrm{OOD}} \sim P_{\mathrm{OOD}}$. Additionally, denote $\mathrm{w}_{\mathrm{spu}}$ as the weights learned by the optimal P_{ID} predictor ($f_{\mathrm{X}}^{P_{\mathrm{ID}}}$) for $Z_{\mathrm{spu}}^{\mathrm{ID}}$. Then, for any $\delta \in (0, 1)$, if

$$\mathrm{w}_{\mathrm{spu}}^\top M \mu_{\mathrm{spu}} + \sqrt{2\kappa^2\, \mathrm{w}_{\mathrm{spu}}^\top M \Sigma_{\mathrm{spu}} M^\top \mathrm{w}_{\mathrm{spu}}\, \log(1/\delta)} < 0, \tag{14}$$

with probability at least $1 - \delta$ over $Z_{\mathrm{spu}}^{\mathrm{OOD}}$, we have

$$\mathrm{acc}_{P_{\mathrm{OOD}}}(f_{\mathrm{X}}^{P_{\mathrm{ID}}}) < \mathrm{acc}_{P_{\mathrm{OOD}}}(f_{\mathrm{dg}}^{\mathcal{E}}), \tag{15}$$

where $f_{\mathrm{dg}}^{\mathcal{E}}$ and $f_{\mathrm{X}}^{P_{\mathrm{ID}}}$ are the optimal domain-general and domain-specific predictions (Definitions 3-4).

Proof. Define $Z_{\mathrm{spu}}^{\mathrm{OOD}} = M Z_{\mathrm{spu}}$. From Equation 1 and the law of total probability, the out-of-distribution (OOD) accuracy of $f_{\mathrm{X}}^{P_{\mathrm{ID}}}$ is:

$$\mathrm{acc}_{\mathrm{OOD}}(f_{\mathrm{X}}^{P_{\mathrm{ID}}}) = \int_{a \in \mathbb{R}} \Pr(\mathrm{w}_{\mathrm{dg}}^\top Z_{\mathrm{dg}} > -a)\, dP_{\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}^{\mathrm{OOD}}}(a),$$

where $P_{\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}^{\mathrm{OOD}}}$ is the distribution of $\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}^{\mathrm{OOD}}$. Additionally:

$$\mathrm{acc}_{\mathrm{OOD}}(f_{\mathrm{dg}}^{\mathcal{E}}) = \Pr(\mathrm{w}_{\mathrm{dg}}^\top Z_{\mathrm{dg}} > 0).$$

It suffices to show:

$$\Pr(\mathrm{w}_{\mathrm{dg}}^\top Z_{\mathrm{dg}} > -\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}^{\mathrm{OOD}}) < \Pr(\mathrm{w}_{\mathrm{dg}}^\top Z_{\mathrm{dg}} > 0) \tag{16}$$

with high probability.

Since Z_{spu} is sub-Gaussian with parameter κ, $\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}^{\mathrm{OOD}} = \mathrm{w}_{\mathrm{spu}}^\top M Z_{\mathrm{spu}}$ is also sub-Gaussian with mean:

$$\mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}^{\mathrm{OOD}}] = \mathrm{w}_{\mathrm{spu}}^\top M \mu_e,$$

and sub-Gaussian parameter $\kappa^2\, \mathrm{w}_{\text{spu}}^\top M\Sigma_{\text{spu}}M^\top \mathrm{w}_{\text{spu}}$ where $\mu_e = \mathbb{E}_{P_{\text{ID}}}[Z_{\text{spu}}]$ and Σ_{spu} is the covariance of Z_{spu}. Thus:

$$\Pr\big(\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}} > \mathbb{E}[\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}}] + t\big) \le \exp\Big(-\frac{t^2}{2\kappa^2\, \mathrm{w}_{\text{spu}}^\top M\Sigma_{\text{spu}}M^\top \mathrm{w}_{\text{spu}}}\Big), \quad \forall\, t > 0.$$

Choose

$$t = \sqrt{2\kappa^2\, (\mathrm{w}_{\text{spu}}^\top M\Sigma_{\text{spu}}M^\top \mathrm{w}_{\text{spu}})\, \log(1/Z_{\text{spu}}^{\text{OOD}})}.$$

Then:

$$\Pr\big(\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}} > \mathrm{w}_{\text{spu}}^\top M\mu_e + t\big) \le Z_{\text{spu}}^{\text{OOD}}.$$

Hence, with probability at least $1 - Z_{\text{spu}}^{\text{OOD}}$:

$$\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}} < \mathrm{w}_{\text{spu}}^\top M\mu_e + \sqrt{2\kappa^2\, \mathrm{w}_{\text{spu}}^\top M\Sigma_{\text{spu}}M^\top \mathrm{w}_{\text{spu}}\, \log(1/Z_{\text{spu}}^{\text{OOD}})}.$$

Assume:

$$\mathrm{w}_{\text{spu}}^\top M\mu_e + \sqrt{2\kappa^2\, (\mathrm{w}_{\text{spu}}^\top M\Sigma_{\text{spu}}M^\top \mathrm{w}_{\text{spu}})\, \log(1/Z_{\text{spu}}^{\text{OOD}})} < 0.$$

Therefore, $\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}} < 0$ with probability at least $1 - Z_{\text{spu}}^{\text{OOD}}$. In this case:

$$\{\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} > -\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}}\} \subset \{\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} > 0\},$$

which implies:

$$\Pr(\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} > -\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}}) < \Pr(\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} > 0).$$

Equivalently:

$$\begin{aligned} \mathrm{acc}_{\text{OOD}}(f_{\mathrm{X}}^{P_{\text{ID}}}) &= \Pr(\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} + \mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}} > 0) \\ &= \Pr(\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} > -\mathrm{w}_{\text{spu}}^\top Z_{\text{spu}}^{\text{OOD}}) \\ &< \Pr(\mathrm{w}_{\text{dg}}^\top Z_{\text{dg}} > 0) \\ &= \mathrm{acc}_{\text{OOD}}(f_{\text{dg}}^{\mathcal{E}}). \end{aligned} \tag{17}$$

Therefore, with probability at least $1 - \delta$:

$$\mathrm{acc}_{\text{OOD}}(f_{\mathrm{X}}^{P_{\text{ID}}}) < \mathrm{acc}_{\text{OOD}}(f_{\text{dg}}^{\mathcal{E}}).$$

□

A.3 Corollary 3—Conditions For Well-Specified Domain Generalization Benchmark Splits For Gaussian Spurious Features

Corollary 3 (Gaussianity Features)**.** *Suppose Z_{spu} is Gaussian, and WLOG $\Sigma_{spu} = \Sigma_{dg} = I_m$, and $\|\mathbb{E}_P[Z_{dg}]\| = 1$.*

$$\max_{f\in\mathcal{F}\setminus\mathcal{F}_{dg}} acc_{OOD}(f_X) < acc_{OOD}(f_{dg}^*)$$

if and only if

- **Spurious Correlation Reversal**
$$\mu_{spu}^T M \mu_{spu} < 0 \tag{18}$$
- **Controlled Spurious Feature OOD Variance**
$$\|\mu_{spu}\|^2 > (\mu_{spu}^T M\mu_{spu})^2 + 2.5(\mu_{spu}^T M\mu_{spu}) \tag{19}$$
For $\lambda_{\max}$ optimistically and $\lambda_{\min}$ pessimistically, one needs the following for the eigenvalues of M:
$$\|\mu_{spu}\|^2 < \frac{1-2.5\lambda_{\min}}{\lambda_{\min}^2} \qquad \|\mu_{spu}\|^2 < \frac{1-2.5\lambda_{\max}}{\lambda_{\max}^2}. \tag{20}$$
Equation 20 does not hold if $\lambda > \frac{1}{2.5}$.

Note that these conditions are necessary and sufficient under Gaussianity.

A.4 Proof of Theorem 2—Conditions for Accuracy on the Line

Lemma 2. *Assume Z_{spu}^{ID} is sub-Gaussian with mean μ_{spu}, variance Σ_{spu}, and parameter κ. Define $M \in \mathbb{R}^{l\times l}$ such that $Z_{spu}^{OOD} = MZ_{spu}$ where in-distribution $Z_{spu}^{OOD} \sim P_{OOD}$ and out-of-distribution $Z_{spu}^{OOD} \sim P_{OOD}$. Additionally, denote w_{spu} as the weights learned by the optimal P_{ID} predictor ($f_X^{P_{ID}}$) for Z_{spu}^{ID}. Further, assume that*

$$\|M\mu_e - \mu_e\| \leq \epsilon_1, \tag{21}$$

$$\left\|w_{spu}^\top M\Sigma_{spu}M^\top w_{spu} - w_{spu}^\top \Sigma_{spu} w_{spu}\right\| \leq \epsilon_2, \tag{22}$$

and there exists a constant $B > 0$ such that for sufficiently small t,

$$\Pr(|f_X(X)| \leq t) \leq Bt. \tag{23}$$

Then, for any $\delta > 0$, with probability at least $1-\delta$ over Z_{spu}, the following holds for any classifiers $f_X \in \mathcal{F}$:

$$|acc_P(f_X) - acc_{P_M}(f_X)| \leq B\cdot\epsilon, \tag{24}$$

where

$$\epsilon = \|w_{spu}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2},$$

and

$$C = \kappa\cdot\max\left\{\|w_{spu}\|, \|M\|\cdot\|w_{spu}\|\right\}.$$

Proof. First, observe that

$$\Delta(X) = f_{\mathrm{X}}(X) - f_{\mathrm{X}}(X') = \mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}} - \mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}} = \mathrm{w}_{\mathrm{spu}}^\top(Z_{\mathrm{spu}} - MZ_{\mathrm{spu}}), \tag{25}$$

where $X \sim P$ and $X' \in P_M$.
We then decompose $\Delta(X)$ into its deterministic part and stochastic part:

$$\Delta(X) = \underbrace{\left[\mathrm{w}_{\mathrm{spu}}^\top\mathbb{E}[Z_{\mathrm{spu}}] - \mathrm{w}_{\mathrm{spu}}^\top\mathbb{E}[MZ_{\mathrm{spu}}]\right]}_{:=g(X)} + \underbrace{(\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}]) - (\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}}])}_{:=h(X)} \tag{26}$$

From the Equation 30, applying the Cauchy-Schwarz inequality:

$$\begin{aligned} g(x) = \left|\mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}}] - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}]\right| &= \left|\mathrm{w}_{\mathrm{spu}}^\top M\mu_e - \mathrm{w}_{\mathrm{spu}}^\top\mu_e\right| \\ &\leq \|\mathrm{w}_{\mathrm{spu}}\|\cdot\|M\mu_e - \mu_e\| \\ &\leq \|\mathrm{w}_{\mathrm{spu}}\|\epsilon_1. \end{aligned}$$

Now we consider $h(X)$. Since Z_{spu} is sub-Gaussian with parameter κ, both Z and Z_M are sub-Gaussian with parameters $\kappa\|\mathrm{w}_{\mathrm{spu}}\|$ and $\kappa\|M\|\|\mathrm{w}_{\mathrm{spu}}\|$, respectively. Therefore, for any $t > 0$:

$$\Pr\left(|\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}]| > t\right) \leq 2\exp\left(-\frac{t^2}{2(\kappa\|\mathrm{w}_{\mathrm{spu}}\|)^2}\right), \tag{27}$$

$$\Pr(|\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}}]\| > t) \leq 2\exp\left(-\frac{t^2}{2(\kappa\|M\|\|\mathrm{w}_{\mathrm{spu}}\|)^2}\right).$$

Applying the union bound, to ensure that both deviations hold simultaneously with probability at least $1-\delta$, choose

$$t = C\sqrt{2\log(4\delta)},$$

where

$$C = \kappa\cdot\max\left\{\|\mathrm{w}_{\mathrm{spu}}\|, \|M\|\cdot\|\mathrm{w}_{\mathrm{spu}}\|\right\}.$$

Thus, with probability at least $1-\delta$:

$$\left|\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}]\right| \leq t \quad \text{and} \quad \left|\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^\top MZ_{\mathrm{spu}}]\right| \leq t.$$

Substituting t, we have with probability at least $1-\delta$:

$$\left|\mathrm{w}_{\mathrm{spu}}^{\top} Z_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^{\top} Z_{\mathrm{spu}}]\right| + \left|\mathrm{w}_{\mathrm{spu}}^{\top} M Z_{\mathrm{spu}} - \mathbb{E}[\mathrm{w}_{\mathrm{spu}}^{\top} M Z_{\mathrm{spu}}]\right| \leq C\sqrt{2\log(4/\delta)} \tag{28}$$

Additionally, Equation 31 gives us a bound on the covariance difference:

$$\left|\mathrm{w}_{\mathrm{spu}}^{\top} M \Sigma_{\mathrm{spu}} M^{\top} \mathrm{w}_{\mathrm{spu}} - \mathrm{w}_{\mathrm{spu}}^{\top} \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}}\right| \leq \epsilon_2$$

Combining Equations 27, 28, 31, we have with probability $1-\delta$,

$$\Delta(X) \leq \|\mathrm{w}_{\mathrm{spu}}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2} = \epsilon. \tag{29}$$

The classifier's accuracy difference is influenced by the probability that $f_{\mathrm{X}}^{P}(X)$ and $f_{\mathrm{X}}^{P_M}(X)$ disagree in sign. Given the bounded difference in outputs and margin condition (Equation 32, this probability is controlled by $B\epsilon$. Therefore, we have

$$|\mathrm{acc}_P(f_{\mathrm{X}}) - \mathrm{acc}_{P_M}(f_{\mathrm{X}})| \leq B\epsilon.$$

This completes the proof. ☐

A.4.1 Theorem 2

Assume $Z_{\mathrm{spu}}^{\mathrm{ID}}$ is sub-Gaussian with mean μ_{spu}, variance Σ_{spu}, and parameter κ. Define $M \in \mathbb{R}^{l\times l}$ such that $Z_{\mathrm{spu}}^{\mathrm{OOD}} = MZ_{\mathrm{spu}}$ where in-distribution $Z_{\mathrm{spu}}^{\mathrm{OOD}} \sim P_{\mathrm{OOD}}$ and out-of-distribution $Z_{\mathrm{spu}}^{\mathrm{OOD}} \sim P_{\mathrm{OOD}}$. Additionally, denote $\mathrm{w}_{\mathrm{spu}}$ as the weights learned by the optimal P_{ID} predictor ($f_{\mathrm{X}}^{P_{\mathrm{ID}}}$) for $Z_{\mathrm{spu}}^{\mathrm{ID}}$. Further, assume that

$$\|M\mu_e - \mu_e\| \leq \epsilon_1, \tag{30}$$

$$\left\|\mathrm{w}_{\mathrm{spu}}^{\top} M \Sigma_{\mathrm{spu}} M^{\top} \mathrm{w}_{\mathrm{spu}} - \mathrm{w}_{\mathrm{spu}}^{\top} \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}}\right\| \leq \epsilon_2,, \tag{31}$$

there exists a constant $B > 0$ such that for sufficiently small t,

$$\Pr(|f_{\mathrm{X}}(X)| \leq t) \leq Bt, \tag{32}$$

and there exists a $\alpha > 0$ such that,

$$\mathrm{acc}_P(f_{\mathrm{X}}) \in [\alpha, 1-\alpha] \quad \text{and} \quad a\mathrm{acc}_{P_M}(f_{\mathrm{X}}) \in [\alpha, 1-\alpha].$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|\Phi^{-1}(\mathrm{acc}_P(f_{\mathrm{X}})) - a\Phi^{-1}(\mathrm{acc}_{P_M}(f_{\mathrm{X}}))| \leq \widetilde{\epsilon}, \tag{33}$$

for any classifier $f_{\mathrm{X}} \in \mathcal{F}$, where

$$\widetilde{\epsilon} = LB\big(\|\mathrm{w}_{\mathrm{spu}}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2}\big) + |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|$$

and Φ is the Gaussian cumulative density function with Lipschitz constant L in $[\alpha, 1-\alpha]$.

Proof. By Lemma 2, with probability at least $1-\delta$ we have

$$\left|\mathrm{acc}_P(f_{\mathrm{X}}) - a\mathrm{acc}_{P_M}(f_{\mathrm{X}})\right| \leq B\big(\|\mathrm{w}_{\mathrm{spu}}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2}\big),$$

where $C = \kappa \max\{\|\mathrm{w}_{\mathrm{spu}}\|, \|M\|\|\mathrm{w}_{\mathrm{spu}}\|\}$. Separately, by assumption $\mathrm{acc}_P(f_{\mathrm{X}}) \in [\alpha, 1-\alpha]$ and $a\mathrm{acc}_{P_M}(f_{\mathrm{X}}) \in [\alpha, 1-\alpha]$, therefore $\Phi^{-1}(\mathrm{acc}_P(f_{\mathrm{X}}))$ and $\Phi^{-1}(a\mathrm{acc}_{P_M}(f_{\mathrm{X}}))$ lie in a region where the standard normal probability density function $\phi(\cdot)$ is bounded away from 0. Recall that $\Phi^{-1}(x)$ is Lipschitz on $[\alpha, 1-\alpha]$. Explicitly, if $p, q \in [\alpha, 1-\alpha]$, then

$$|\Phi^{-1}(p) - \Phi^{-1}(q)| \leq L|p-q|,$$

where

$$L = \sup_{x\in[\alpha,1-\alpha]} \left|\frac{d}{dx}\Phi^{-1}(x)\right| = \sup_{u\in[\Phi^{-1}(\alpha),\Phi^{-1}(1-\alpha)]} \frac{1}{\phi(u)} < \infty.$$

Let $p = \mathrm{acc}_P(f_X)$ and $q = \mathrm{acc}_{P_M}(f_X)$. Then

$$|\Phi^{-1}(p) - a\Phi^{-1}(q)| \leq |\Phi^{-1}(p) - \Phi^{-1}(q)| + |1-a||\Phi^{-1}(q)| \tag{34}$$
$$\leq L|p-q| + |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|. \tag{35}$$

Then, $|p-q| \leq C\sqrt{\log(1/\delta)}(\epsilon_1 + \sqrt{\epsilon_2})$. Thus we obtain

$$|\Phi^{-1}(\mathrm{acc}_P(f_X)) - a\Phi^{-1}(\mathrm{acc}_{P_M}(f_X))| \leq LB\big(\|\mathrm{w}_{\mathrm{spu}}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2}\big) + |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|.$$

We denote the sum of these terms by $\widetilde{\epsilon}$. Hence Equation 33 holds with probability at least $1-\delta$. □

A.5 Proof of Corollary 1—Tradeoff Between Accuracy on The Line and Well-Specification

Let $Z_{\mathrm{spu}} \in \mathbb{R}^k$ be sub-Gaussian with parameter κ and covariance Σ_{spu}. Fix $\mathrm{w}_{\mathrm{spu}} \in \mathbb{R}^k$ such that $\mathrm{w}_{\mathrm{spu}}^\top \mu_e > 0$. Suppose $M \in \mathbb{R}^{k\times k}$ satisfies *(spurious correlation reversal condition)*

$$\mathrm{w}_{\mathrm{spu}}^\top(M\mu_e) + \sqrt{2\kappa^2 \mathrm{w}_{\mathrm{spu}}^\top M\Sigma_{\mathrm{spu}}M^\top \mathrm{w}_{\mathrm{spu}} \log(1/\delta)} \leq -\gamma < 0 \tag{36}$$

for some margin $\gamma > 0$.

Assume also that there exists some $\alpha > 0$ such that,

$$\mathrm{acc}_P(f_X),\, a\cdot \mathrm{acc}_{P_M}(f_X) \in [\alpha, 1-\alpha].$$

Then with probability at least $1-\delta$,

$$|\Phi^{-1}(\mathrm{acc}_P(f_X)) - a\Phi^{-1}(\mathrm{acc}_{P_M}(f_X))| \geq C\kappa\|\mathrm{w}_{\mathrm{spu}}\|\sqrt{\log(1/\delta)}\|M\mu_e - \mu_e\| - \zeta, \tag{37}$$

where $\zeta = |1-a|\max_{x\in[\alpha,1-\alpha]}|\Phi^{-1}(x)|$, for some positive constant C (depending on α and the local slope of Φ^{-1}), and Φ is the Gaussian cumulative density function. Moreover,

$$\|M\mu_e - \mu_e\| \geq \|\mathrm{w}_{\mathrm{spu}}\|^{-1}(\gamma + \mathrm{w}_{\mathrm{spu}}^\top \mu_e),$$

so the right side of 37 is strictly positive whenever $\gamma + \mathrm{w}_{\mathrm{spu}}^\top\mu_e > 0$.

Proof. Since $\mathrm{w}_{\mathrm{spu}}^\top(M\mu_e)$ is sufficiently negative (compared to the random variations of $\mathrm{w}_{\mathrm{spu}}^\top(MZ_{\mathrm{spu}})$) by 36, it must also be that

$$\mathrm{w}_{\mathrm{spu}}^\top(M\mu_e) \leq -\gamma < 0.$$

Meanwhile, $\mathrm{w}_{\mathrm{spu}}^\top\mu_e > 0$. Combining these gives

$$\mathrm{w}_{\mathrm{spu}}^\top(M\mu_e - \mu_e) = \mathrm{w}_{\mathrm{spu}}^\top(M\mu_e) - \mathrm{w}_{\mathrm{spu}}^\top(\mu_e) \leq -\gamma - \mathrm{w}_{\mathrm{spu}}^\top(\mu_e) = -(\gamma + \mathrm{w}_{\mathrm{spu}}^\top\mu_e). \tag{38}$$

By Cauchy–Schwarz,

$$\|\mathrm{w}_{\mathrm{spu}}\|\|M\mu_e - \mu_e\| \geq |\mathrm{w}_{\mathrm{spu}}^\top(M\mu_e - \mu_e)| \geq \gamma + \mathrm{w}_{\mathrm{spu}}^\top\mu_e > 0.$$

Hence

$$\|M\mu_e - \mu_e\| \geq \|\mathrm{w}_{\mathrm{spu}}\|^{-1}(\gamma + \mathrm{w}_{\mathrm{spu}}^\top\mu_e). \tag{39}$$

Note

$$\mathrm{acc}_P(f_X) = \Pr(\mathrm{w}_{\mathrm{dg}}^\top Z_{\mathrm{dg}} > -\mathrm{w}_{\mathrm{spu}}^\top Z_{\mathrm{spu}}), \quad \mathrm{acc}_{P_M}(f_X) = \Pr(\mathrm{w}_{\mathrm{dg}}^\top Z_{\mathrm{dg}} > -\mathrm{w}_{\mathrm{spu}}^\top(MZ_{\mathrm{spu}})).$$

Because $\mathrm{w}_{\mathrm{spu}}^\top(M\mu_e)$ is sufficiently negative (by 36) and $\mathrm{w}_{\mathrm{spu}}^\top\mu_e$ is positive, one can apply standard sub-Gaussian concentration inequality to $\mathrm{w}_{\mathrm{spu}}^\top(Z_{\mathrm{spu}} - \mu_e)$ and $\mathrm{w}_{\mathrm{spu}}^\top(MZ_{\mathrm{spu}} - M\mu_e)$ to show that with high probability (at least $1-\delta$),

$$|\mathrm{acc}_P(f_\mathrm{X}) - a\mathrm{acc}_{P_M}(f_\mathrm{X})| \geq C_0\kappa\|\mathrm{w}_\mathrm{spu}\|\sqrt{\log(1/\delta)}\|M\mu_e - \mu_e\|,$$

for some constant $C_0 > 0$.

In other words, the spurious correlation reversal ensures the probability under MZ_spu is drastically smaller or larger—yielding a bigger lower bound on $|\mathrm{acc}_P(f_\mathrm{X}) - a \cdot \mathrm{acc}_{P_M}(f_\mathrm{X})|$.

By assumption, we have $\mathrm{acc}_P(f_\mathrm{X})$, $a \cdot \mathrm{acc}_{P_M}(f_\mathrm{X}) \in [\alpha, 1-\alpha]$. Recall that on $[\alpha, 1-\alpha]$, Φ^{-1} is Lipschitz, i.e.,

$$|\Phi^{-1}(p) - \Phi^{-1}(q)| \geq L|p-q| \quad \text{for all } p, q \in [\alpha, 1-\alpha].$$

Thus, with probability at least $1-\delta$,

$$|\Phi^{-1}(\mathrm{acc}_P(f_\mathrm{X})) - a \cdot \Phi^{-1}(\mathrm{acc}_{P_M}(f_\mathrm{X}))| \geq L|\mathrm{acc}_P(f_\mathrm{X}) - a\mathrm{acc}_{P_M}(f_\mathrm{X})| - |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)| \tag{40}$$

$$\geq L(C_0\kappa\|\mathrm{w}_\mathrm{spu}\|\sqrt{\log(1/\delta)}\|M\mu_e - \mu_e\|) - |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)| \tag{41}$$

$$= (C_0L)\kappa\|\mathrm{w}_\mathrm{spu}\|\sqrt{\log(1/\delta)}\|M\mu_e - \mu_e\| - |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|. \tag{42}$$

Hence, the strength of *accuracy on the line* is also lower bounded by a constant multiple of $\|M\mu_e - \mu_e\| \neq 0$. This completes the proof. ☐

A.6 Proof of Corollary 2—Shifts with perfect accuracy on the line are misspecified almost everywhere

Define

$$\mathcal{W}_\epsilon = \left\{ M \in \mathbb{R}^{k\times k} : \begin{array}{l} \mathrm{w}_\mathrm{spu}^\top M\mu_e + \sqrt{2\kappa^2\, \mathrm{w}_\mathrm{spu}^\top M\Sigma_e M^\top \mathrm{w}_\mathrm{spu}\, \log(1/\delta)} < 0, \\ \left|\Phi^{-1}(\mathrm{acc}_P(f_\mathrm{X})) - a\,\Phi^{-1}(\mathrm{acc}_{P_M}(f_\mathrm{X}))\right| \leq \epsilon \end{array} \right\}.$$

Then:

(i) $\mathcal{W}_0$ has Lebesgue measure zero in $\mathbb{R}^{k\times k}$.

(ii) For any $0 \leq \epsilon_i \leq \epsilon_j$, we have $\mathcal{W}_{\epsilon_i} \subseteq \mathcal{W}_{\epsilon_j}$.

In particular, as $\epsilon \to 0$ (i.e., perfect accuracy on the line), almost every shift is misspecified, and the Lebesgue measure of the set of well–specified shifts grows monotonically with ϵ.

Proof. From our derivation (see, e.g., the inequality below)

$$\left|\Phi^{-1}(\mathrm{acc}_P(f_\mathrm{X})) - a\,\Phi^{-1}(\mathrm{acc}_{P_M}(f_\mathrm{X}))\right| \geq L\left(C_0\kappa\|\mathrm{w}_\mathrm{spu}\|\sqrt{\log(1/\delta)}\,\|M\mu_e - \mu_e\|\right) - |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|, \tag{43}$$

if

$$\left|\Phi^{-1}(\mathrm{acc}_P(f_\mathrm{X})) - a\,\Phi^{-1}(\mathrm{acc}_{P_M}(f_\mathrm{X}))\right| \leq \epsilon,$$

then for ϵ sufficiently small it must be that

$$\|M\mu_e - \mu_e\| \leq \frac{\epsilon + |1-a| \max_{x\in[\alpha,1-\alpha]} |\Phi^{-1}(x)|}{L\,C_0\kappa\|\mathrm{w}_\mathrm{spu}\|\sqrt{\log(1/\delta)}}.$$

Thus, as $\epsilon \to 0$, we must have

$$\|M\mu_e - \mu_e\| = 0,$$

i.e.,

$$M\mu_e = \mu_e.$$

Let

$$S = \{M \in \mathbb{R}^{k\times k} : M\mu_e = \mu_e\}.$$

Since $\mu_e \neq 0$, S is an affine subspace of of $\mathbb{R}^{k\times k}$ with dimension lower than k^2 and thus has Lebesgue measure zero. Since,

$$\mathcal{W}_0 \subset S,$$

and $\mathcal{W}_0$ has Lebesgue measure zero.

The monotonicity claim is follows immediately: if $0 \leq \epsilon_i \leq \epsilon_j$, then by definition

$$\mathcal{W}_{\epsilon_i} \subseteq \mathcal{W}_{\epsilon_j}.$$

This completes the proof. $\square$

A.7 Example of Shifts with Accuracy on the Line that are Well-Specified

Let $Z_{\text{spu}} \in \mathbb{R}^k$ be sub-Gaussian with parameter κ, mean $\mu_e = \mathbb{E}[Z_{\text{spu}}] \neq 0$, and covariance Σ_{spu}. Fix $\mathrm{w}_{\text{spu}} \in \mathbb{R}^k$ with $\mathrm{w}_{\text{spu}}^\top \mu_e \neq 0$ and let $a > 0$. Then for any $\epsilon > 0$ and $\delta \in (0,1)$, there exists a matrix $M \in \mathbb{R}^{k\times k}$ such that:

$$\mathrm{w}_{\text{spu}}^\top M\mu_e + \sqrt{2\kappa^2 \mathrm{w}_{\text{spu}}^\top M\Sigma_{\text{spu}} M^\top \mathrm{w}_{\text{spu}} \log(1/\delta)} < 0, \tag{44}$$

$$|\Phi^{-1}(\text{acc}_P(f_{\mathrm{X}})) - a\Phi^{-1}(\text{acc}_{P_M}(f_{\mathrm{X}}))| \leq \epsilon \tag{45}$$

with probability at least $1-\delta$, where $\text{acc}_P(f_{\mathrm{X}})$ and $\text{acc}_{P_M}(f_{\mathrm{X}})$ are defined as in Theorem 2.

Let $v = \frac{\mathrm{w}_{\text{spu}}}{\|\mathrm{w}_{\text{spu}}\|}$ be the unit vector in the direction of w_{spu}, and define its reflection matrix

$$R = I - 2vv^\top.$$

Note that $Rv = -v$ and $R^2 = I$. We then choose a scalar $\alpha > 0$ and define

$$M = \alpha R.$$

Hence

$$\mathrm{w}_{\text{spu}}^\top (M\mu_e) = \mathrm{w}_{\text{spu}}^\top (\alpha R\mu_e) = \alpha(\mathrm{w}_{\text{spu}}^\top R\mu_e) = -\alpha \mathrm{w}_{\text{spu}}^\top \mu_e,$$

$$\mathrm{w}_{\text{spu}}^\top (M\Sigma_{\text{spu}} M^\top)\mathrm{w}_{\text{spu}} = \alpha^2 \mathrm{w}_{\text{spu}}^\top (R\Sigma_{\text{spu}} R^\top)\mathrm{w}_{\text{spu}} = \alpha^2 \mathrm{w}_{\text{spu}}^\top \Sigma_{\text{spu}} \mathrm{w}_{\text{spu}},$$

since $R\mathrm{w}_{\text{spu}} = -\mathrm{w}_{\text{spu}}$ and R is orthogonal.
We now want

$$\mathrm{w}_{\text{spu}}^\top (M\mu_e) + \sqrt{2\kappa^2 \mathrm{w}_{\text{spu}}^\top (M\Sigma_{\text{spu}} M^\top)\mathrm{w}_{\text{spu}} \log(1/\delta)} < 0.$$

Since $\mathrm{w}_{\text{spu}}^\top (M\mu_e) = -\alpha \mathrm{w}_{\text{spu}}^\top \mu_e$ and $\mathrm{w}_{\text{spu}}^\top (M\Sigma_{\text{spu}} M^\top)\mathrm{w}_{\text{spu}} = \alpha^2 \mathrm{w}_{\text{spu}}^\top \Sigma_{\text{spu}} \mathrm{w}_{\text{spu}}$, this inequality becomes

$$-\alpha|\mathrm{w}_{\text{spu}}^\top \mu_e| + \alpha\sqrt{2\kappa^2 \mathrm{w}_{\text{spu}}^\top \Sigma_{\text{spu}} \mathrm{w}_{\text{spu}} \log(1/\delta)} < 0,$$

i.e.

$$\alpha\left(-|\mathrm{w}_{\text{spu}}^\top \mu_e| + \sqrt{2\kappa^2 \mathrm{w}_{\text{spu}}^\top \Sigma_{\text{spu}} \mathrm{w}_{\text{spu}} \log(1/\delta)}\right) < 0.$$

Hence we can choose α such that

$$\alpha > \frac{\sqrt{2\kappa^2 \mathrm{w}_{\mathrm{spu}}^\top \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}} \log(1/\delta)}}{|\mathrm{w}_{\mathrm{spu}}^\top \mu_e|},$$

so that the negative term $-\alpha|\mathrm{w}_{\mathrm{spu}}^\top \mu_e|$ dominates.
Then we have, from Theorem 2, that

$$|\Phi^{-1}(\mathrm{acc}_P(f_{\mathrm{X}})) - a\Phi^{-1}(\mathrm{acc}_{P_M}(f_{\mathrm{X}}))| \leq B\big(\|\mathrm{w}_{\mathrm{spu}}\|\epsilon_1 + C\sqrt{2\log(4/\delta)} + \sqrt{\epsilon_2}\big)$$

with probability at least $1-\delta$, where

- $C = \kappa \max\{\|\mathrm{w}_{\mathrm{spu}}\|, \|M\|\|\mathrm{w}_{\mathrm{spu}}\|\}$ (up to a small Lipschitz factor if $a \neq 1$),
- $\epsilon_1 = \|M\mu_e - \mu_e\|$,
- $\epsilon_2 = |\mathrm{w}_{\mathrm{spu}}^\top(\Sigma_{\mathrm{spu}} - M\Sigma_{\mathrm{spu}}M^\top)\mathrm{w}_{\mathrm{spu}}|$.

$$\epsilon_1 = \|M\mu_{\mathrm{spu}} - \mu_{\mathrm{spu}}\| = \|\alpha R\mu_{\mathrm{spu}} - \mu_{\mathrm{spu}}\|,$$

and

$$\epsilon_2 = |\mathrm{w}_{\mathrm{spu}}^\top \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}} - \alpha^2 \mathrm{w}_{\mathrm{spu}}^\top \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}}| = |\alpha^2 - 1| \cdot |\mathrm{w}_{\mathrm{spu}}^\top \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}}|,$$

We want to mane ϵ_1, ϵ_2 arbitrarily small while also keeping α large enough to satisfy the spurious correlation reversal condition.

Thus, we define

$$\alpha = \max\left\{1+\eta, \frac{\sqrt{2\kappa^2 \mathrm{w}_{\mathrm{spu}}^\top \Sigma_{\mathrm{spu}} \mathrm{w}_{\mathrm{spu}} \log(1/\delta)}}{|\mathrm{w}_{\mathrm{spu}}^\top \mu_e|}\right\}$$

for some small $\eta > 0$ chosen so that $|\alpha - 1| < \delta'$ implies ϵ_2 (and ϵ_1) are each below the desired threshold that ensures the final bound is $< \epsilon$.

By choosing α accordingly, (1) We overpower $|\mathrm{w}_{\mathrm{spu}}^\top \mu_e|$ in the spurious correlation reversal term, thus guaranteeing 44, and (2) we keep α sufficiently close to 1 so that ϵ_1 and $|\alpha^2 - 1|$ remain small, thereby making $(\epsilon_1 + \sqrt{\epsilon_2})$ small enough that the Theorem 2 bound is at most ϵ.

Hence $M = \alpha R$ satisfies both 44 and 45 with probability at least $1-\delta$. Note, however, that the family of such M has a Lebesgue measure of zero.

A.8 Finite Mixtures of Sub-Gaussians are Sub-Gaussian

Lemma 3. *Let X be a finite mixture of sub-Gaussian random variables $X_1, X_2, \ldots, X_k$ with parameters $c_1, c_2, \ldots, c_k$ respectively. That is, for all $t \in \mathbb{R}$ and each $i \in \{1, \ldots, k\}$,*

$$\mathbb{E}\big[e^{t(X_i - \mathbb{E}[X_i])}\big] \leq e^{c_i t^2}.$$

Assume the mixture probabilities $p_1, p_2, \ldots, p_k$ satisfy $\sum_{i=1}^k p_i = 1$ and $p_i \geq 0$. Then X is also subgaussian. Specifically, there exists a constant $c > 0$ such that for all $t \in \mathbb{R}$,

$$\mathbb{E}\big[e^{t(X - \mathbb{E}[X])}\big] \leq e^{ct^2}.$$

Proof. Since X is a mixture, we have

$$\mathbb{E}\big[e^{t(X-\mathbb{E}[X])}\big] = \sum_{i=1}^{k} p_i \, \mathbb{E}\big[e^{t(X_i-\mathbb{E}[X])}\big].$$

For each i, write

$$X_i - \mathbb{E}[X] = \big(X_i - \mathbb{E}[X_i]\big) + \big(\mathbb{E}[X_i] - \mathbb{E}[X]\big).$$

Thus,

$$\mathbb{E}\big[e^{t(X_i-\mathbb{E}[X])}\big] = e^{t(\mathbb{E}[X_i]-\mathbb{E}[X])} \, \mathbb{E}\big[e^{t(X_i-\mathbb{E}[X_i])}\big] \le e^{t(\mathbb{E}[X_i]-\mathbb{E}[X])} \, e^{c_i \, t^2}.$$

Let

$$\Delta = \max_{1\le i\le k} \big|\mathbb{E}[X_i] - \mathbb{E}[X]\big| \quad \text{and} \quad C = \max_{1\le i\le k} c_i.$$

Then, since $e^{t(\mathbb{E}[X_i]-\mathbb{E}[X])} \le e^{|t|\Delta}$ for each i, it follows that

$$\mathbb{E}\big[e^{t(X-\mathbb{E}[X])}\big] \le \sum_{i=1}^{k} p_i \, e^{|t|\Delta} \, e^{C\,t^2} = e^{|t|\Delta} \, e^{C\,t^2}.$$

Then we have for all $t \in \mathbb{R}$

$$e^{|t|\Delta} \le e^{\frac{1}{2}\Delta^2} e^{\frac{1}{2}t^2}$$

and

$$\mathbb{E}\big[e^{t(X-\mathbb{E}[X])}\big] \le e^{\frac{1}{2}\Delta^2} \, e^{\left(C+\frac{1}{2}\right)t^2}.$$

Defining

$$c = C + \frac{1}{2} + \frac{\frac{1}{2}\Delta^2}{t^2},$$

note that the factor $e^{\frac{1}{2}\Delta^2}$ is independent of t and can be absorbed into a constant. In particular, there exists a constant $c' > 0$ (which may depend on Δ and C) such that

$$\mathbb{E}\big[e^{t(X-\mathbb{E}[X])}\big] \le e^{c'\,t^2} \quad \text{for all } t \in \mathbb{R}.$$

Thus, X is subgaussian. □

B Simulation Experiment Setup

Simulation Experiments. We evaluate our results so far empirically. We define an initial distribution with $Z_{\text{dg}} \in \mathbb{R}^2$ as a Gaussian with mean $Y \cdot \mu_{\text{dg}}$, where $\mu_{\text{dg}} = [1; 1]$ and unit variance, and $Z^{\text{ID}}_{\text{spu}} \in \mathbb{R}^2$ as a Gaussian with mean $Y \cdot \mu_{\text{spu}}$, where $\mu_{\text{spu}} = [1; 1]$ and unit variance. The input $X \in \mathbb{R}^4$ and label $y \in \{0, 1\}$. We define a domain by M where $Z^{\text{OOD}}_{\text{spu}} = M Z^{0}_{\text{spu}}$ and all other random variables' distribution is preserved. We consider settings where the training domain is (i) Gaussian and (ii) Sub-Gaussian (mixture of Gausians). We define a set of 50 Gaussian test domains defined by randomly sampled M.

We train two types of models: *domain general*, which are logistic regression models trained and evaluated with only Z_{dg} features but still trained only on the training distribution, and *domain specific*, which are logistic regression models trained an evaluated with X features but still trained only on the training distribution. Details on the experiments can be found in Appendix B.

Figure 1a demonstrates the setting where the training domain is defined by $M = I_{[2]}$, i.e., $Z^{\text{ID}}_{\text{spu}}$ is a multivariate Gaussian. In this setting, we observe the expected behavior derived in Theorem 1. That is, when the spurious correlation reversal and controlled spurious feature variance hold out-of-distribution, the domain-general models outperform the domain-specific models.

Figure 1 demonstrates the setting where the training domain is a mixture of M's, i.e., a mixture of Gaussians making a Sub-Gaussian distribution. Figure 1c demonstrates the setting where M is unconstrained. Here, the test domains can be written as an interpolation of the training domains, i.e., there are positive and negative

Algorithm 1: Generative Mechanism for ColoredMNIST

Input : MNIST dataset with grayscale images z_{dg} and binary labels $y \in \{0, 1\}$
Output : ColoredMNIST dataset with colorized images x and labels y
Define color mapping probability $P(z_{\mathrm{spu}}|y)$ based on a chosen spurious correlation
Sample $y \sim P(y)$ from the original MNIST dataset
Sample grayscale image z_{dg} corresponding to y
`// Introduce spurious correlation`
With probability p, assign color z_{spu} based on $P(z_{\mathrm{spu}}|y)$
With probability $1 - p$, assign color z_{spu} randomly (breaking correlation)
Apply color transformation $T(z_{\mathrm{dg}}, z_{\mathrm{spu}})$ to obtain x
return (x, y)

definite M's mixed to create the training domain. In this setting, Rosenfeld et al. Rosenfeld et al. (2022b) show that the training domain empirical risk minimizer solves the worst-case domain generalization problem. Indeed, figure 1c's results show that there is not generally a difference in OOD performance between the domain-general and domain-specific models.

Figure 1c demonstrates the setting where the testing domains are not a convex combination of the training domain – test domains can be outside the bounds of the training M's. Here, there is a clear difference in the OOD performance between the domain-general and domain-specific models. Furthermore, the expected conditions derived in Theorem 1 are observed. That is, when the spurious correlation reversal and controlled spurious feature variance hold out-of-distribution, the domain-general models outperform the domain-specific models.

Clearly, in natural datasets, it is often impractical to conduct such experiments to determine when a domain-general model achieves the best transfer accuracy on a benchmark. Typically, the domain-general features are unknown, and we lack the ability to manipulate natural datasets. However, we demonstrate below that the absence of a strong positive correlation between in- and out-of-distribution accuracy for arbitrary predictors—referred to as accuracy on the line (Miller et al., 2021), Definition 6—can identify well-specified benchmarks that reliably evaluate domain generalization via transfer accuracy. *We will show that well-specified domain generalization benchmarks exhibit either weak in- and out-of-distribution accuracy correlation or a strong inverse correlation.*

Parameters. We use the following parameters across our experiments: $\mu = [1, 1]$, $\Sigma_{\mathrm{dg}} = \mathrm{diag}([1, 1])$, $\mu_{\mathrm{spu}} = [1, 1]$. We expect our results to hold independent of these parameters. We chose these parameters for the ease of intuition of the results on the simulated dataset. We use a sample size of 1000 for each domain.

$$P_M = \begin{cases} Y & = \mathrm{Bern}(0.5) \\ Z_{\mathrm{dg}} & = \mathcal{N}\big(Y \cdot \mu_{\mathrm{dg}}, \Sigma_{\mathrm{dg}}\big) \\ Z_{\mathrm{spu}} & = \mathcal{N}\big(Y \cdot M\mu_{\mathrm{dg}}, M\Sigma_{\mathrm{spu}}M^{\top}\big) \end{cases} \tag{46}$$

In Figure 1a, We pick a P_I as our train domain and then randomly sample M's to construct P_M's. We then train a domain-specific logistic regression model for $f_{\mathrm{X}} : \mathcal{X} \mapsto \mathcal{Y}$ and a domain-general logistic regression model for $f_{\mathrm{dg}} : \mathcal{Z}_{\mathrm{dg}} \mapsto \mathcal{Y}$ for P_I. We retrain each model

B.1 ColoredMNIST Case Study

The ColoredMNIST dataset (Arjovsky et al., 2019) intuitively illustrates the complexity of benchmarking domain generalization. ColoredMNIST modifies the gray-scale MNIST (Deng, 2012) dataset by adding color as a spurious correlation. The digits labels are binary with +1 when 'digit $\geq$ 5' and -1 otherwise. The observed (training) labels, however, contain 25% label noise, i.e., a predictor that uses digit information can achieve 75% accuracy at most, in/out-of-distribution. Additionally, the digit images are colored. The color of

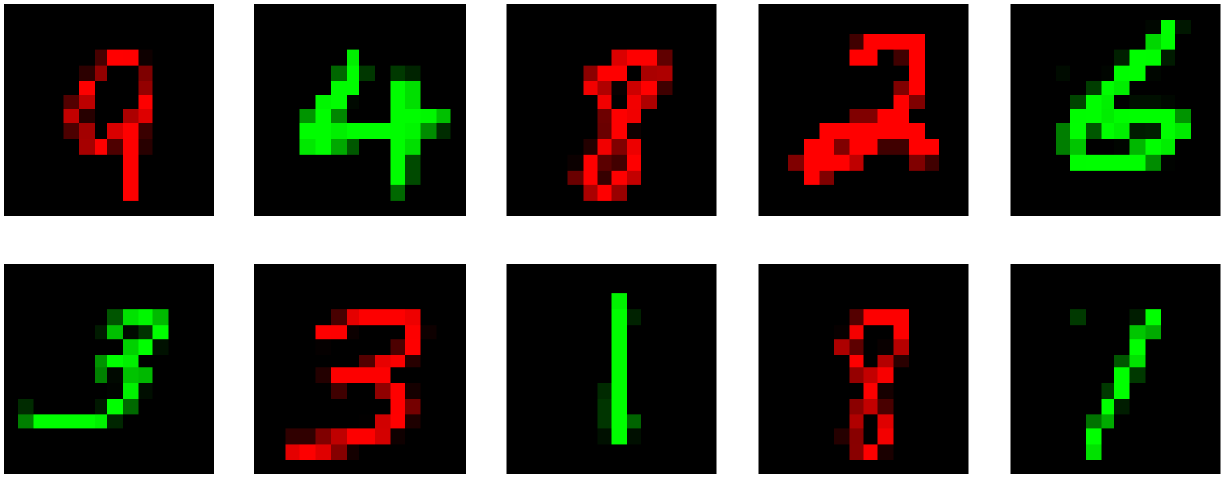
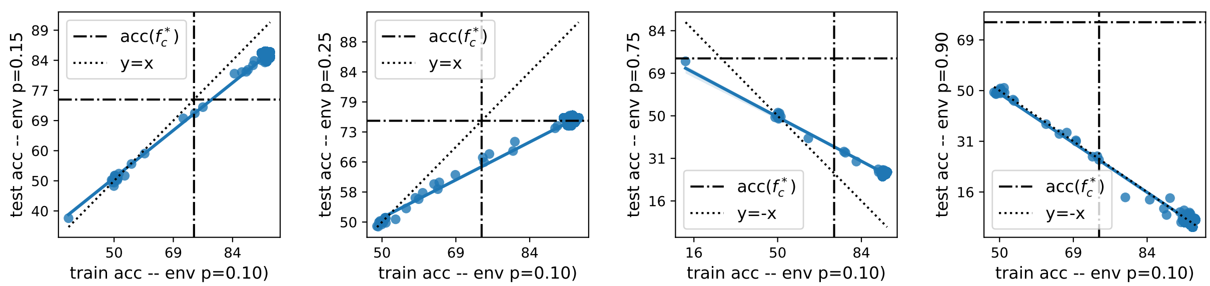


Figure 4: Colored MNIST image examples



(a) m = 0.79
Pearson R = 0.97

(b) m = 0.50
Pearson R = 0.97

(c) m = -0.52
Pearson R = -0.96

(d) m = -0.98
Pearson R = -0.98

Figure 5: Correlations between model performance In-Distristribution vs. Out-of-Distribution on ColoredMNIST variations. m is the slope of the line, and R is the Pearson correlation coefficient. The axis-parallel dashed lines denote the maximum within-domain accuracy of 75%, and $y = x$ represents invariant performance across training and test (target) domains. Models achieving above 75% accuracy use color as a predictor. Figures 5a and 5b represent shifts where color-based predictors achieve the highest OOD accuracy—above 75% accuracy. Without domain knowledge, one might conclude that the best ERM solution is the most domain-general. However, Figures 5c and 5d show that these models are not domain-general; some features that improve ID accuracy hurt OOD performance.

the digit matches the noisy observed labels with probability p_e, inducing a *spurious correlation* or *shortcut* of strength p_e. p_e defines a distinct distribution.

Since the observed labels are noisy versions of the true digit labels, the color potentially correlates more with the observed labels than the digit itself. For example, consider a training domain where $p_e^i = P_i(Y = +1 \mid \text{color} = +1) = 0.9$. A color-based predictor would achieve 90% accuracy in-domain, while a domain-general predictor that ignores color would achieve 75% accuracy at a maximum. Furthermore, under a shift where $p_e^j = P_j(Y = 1 \mid \text{color} = +1) > 0.75$, a color-based model trained on p_e^i model will still outperform the domain-general model in OOD accuracy. However, when $p_e^j = P_j(Y = 1 \mid \text{color} = +1) < 0.75$, the same color-based model will transfer worse than the domain-general model. This simple example underscores that for a domain generalization benchmark to be well-specified, w.r.t. to OOD accuracy, spurious correlations from training to test domains must change enough for the domain-general model to achieve the highest possible OOD accuracy.

Figure 5 demonstrates that domain-general models need not transfer the best OOD. To demonstrate this, we test a set of models on a ColoredMNIST training domain where $P_{\text{tr}}(Y = 1 \mid \text{color} = \text{green}) = 0.1$ and across various test domains with $P_{\text{te}}(Y = 1 \mid \text{color} = \text{green}) = p_e^j$. The observation of the variance in the domain-general and color-based model transfer gap in Figure 5 underscores this work's key question on which ID-OOD shifts allow for reliable domain-generalization evaluation.

We leverage a ConvNet architecture for the ColoredMNIST dataset (Table 3); we vary hyperparameters enumerated in (Gulrajani & Lopez-Paz, 2020a). We vary the hyperparameters in Table 2 and whether or not we use data augmentation.

C Additional Results and Discussion

C.1 Model Training

Data Augmentation. When data augmentation is applied, the transformation consists of a series of preprocessing steps applied to images before they are used for training. First, the image undergoes a **random resized crop** to a size of 224×224 pixels, with a scaling factor ranging from 70% to 100% of the original size. Next, a **random horizontal flip** is applied to introduce variability in orientation. The transformation also includes **color jittering**, which adjusts brightness, contrast, saturation, and hue with a factor of 0.3 each, followed by **random grayscale conversion**, which randomly turns images into grayscale with a certain probability.

Experimental Setup. We follow the following general experimental procedure. When experiments deviate from this, it is specified in their respective sections.

Each dataset consists of E domains, each corresponding to a unique data distribution. Our experiments involve ID/OOD splits using a leave-one-domain-out approach. Specifically, for each domain indexed as $i \in [1..E]$, we train on the subset $\mathcal{E}^i_{\text{train}} = \{e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_E\}$ and test on the held-out domain $\mathcal{E}^i_{\text{test}} = \{e_i\}$.

For each i, we train the following models on $P^{\mathcal{E}^i_{\text{train}}}$: ResNet18, ResNet50, DenseNet121, and ConvNeXt_Tiny. For each model, we consider ImageNet pretrained variants: (i) Fine-tuned – end-to-end training on $P^{\mathcal{E}^i_{\text{train}}}$ and (ii) Transfer learning – retraining only the last layer on $P^{\mathcal{E}^i_{\text{train}}}$. Models are generated with hyperparameters in Table 2; we also take models at different checkpoints during training.

Table 2: Models Generation

Hyperparameter	**Range**
Learning Rate (lr)	10^{-5} to $10^{-3.5}$
Weight Decay	10^{-6} to 10^{-2}
Batch Size	2^3 (8) to $2^{5.5}$ (≈ 45)
Data Augmentation	{True, False}
Transfer Learning	{True, False}
Model Architecture	{ResNet18, ResNet50, DenseNet121, ViT-B-16, and ConvNeXt_Tiny}
Dropout	$\{0.0, 0.1, 0.5\}$
Epoch	–

C.2 ColoredMNIST

ColoredMNIST (Arjovsky et al., 2019). A variant of the MNIST handwritten digit classification dataset (LeCun, 1998). Domain $d \in \{0.1, 0.2, 0.9\}$ contains a disjoint set of digits colored either red or blue.

The label is a noisy function of the digit and color, such that color bears a correlation of d with the label and the digit bears a correlation of 0.75 with the label. This dataset contains 70,000 examples of dimension (2, 28, 28) and 2 classes.

Experimental Details. We leverage a ConvNet architecture for the ColoredMNIST dataset (Table 3).

Table 3: MNIST ConvNet architecture.

#	Layer
1	Conv2D (in=d, out=64)
2	ReLU
3	GroupNorm (groups=8)
4	Conv2D (in=64, out=128, stride=2)
5	ReLU
6	GroupNorm (groups=8)
7	Conv2D (in=128, out=128)
8	ReLU
9	GroupNorm (groups=8)
10	Conv2D (in=128, out=128)
11	ReLU
12	GroupNorm (8 groups)
13	Global average-pooling

Discussion. Despite colored MNIST's apparent simplicity, the spurious correlation between color and the label is quite strong – particularly generalization to text environment 2, going from domains with spurious correlation probability of $0.1, 0.2 \rightarrow 0.9$. in Gulrajani & Lopez-Paz (2020a)'s evaluation of standard domain generalization methods at the time, they found that no model could mitigate the effect of this spurious correlation. We note that this ID/OOD split has a strong accuracy on the inverse line. In test environment 1, we observe that the training distributions are such that the spurious correlations cancel out (0.1 vs. 0.9), and the domain-general model is also the best ID empirical risk minimizer.

Knowledge of the spurious correlation mechanism in each domain makes it relatively easy to identify the type of features a model uses due to the predictability of expected accuracy between models that use color and those that don't. Due to the potential ambiguity of benchmarking results when spurious correlation mechanisms are unknown, semisynthetic benchmarks are vital in the evaluation process.

Table 4: ColoredMNIST ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	1.90	-0.58	0.82	0.00	0.01
Env 1 acc	0.96	0.01	0.94	0.00	0.00
Env 2 acc	-1.56	0.47	-0.74	0.00	0.01

Table 5: ColoredMNIST ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	1.23	-0.12	0.99	0.00	0.00
Env 0 acc	Env 2 acc	-0.73	0.87	-0.36	0.00	0.02
Env 1 acc	Env 0 acc	0.91	0.04	0.98	0.00	0.00
Env 1 acc	Env 2 acc	0.67	0.16	0.72	0.00	0.01
Env 2 acc	Env 0 acc	-1.34	0.53	-0.82	0.00	0.01
Env 2 acc	Env 1 acc	-1.65	0.29	-0.64	0.00	0.02

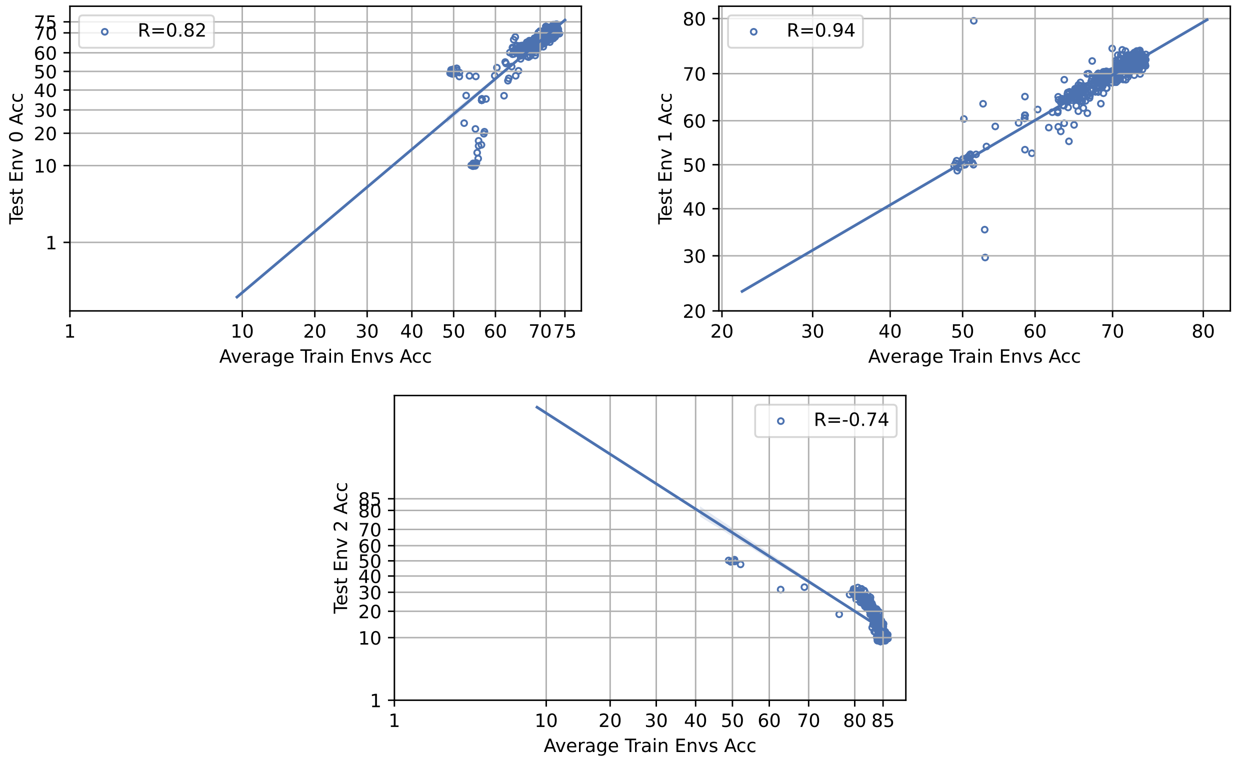


Figure 6: ColoredMNIST: Average train Env Accuracy vs. Test Env Accuracy.

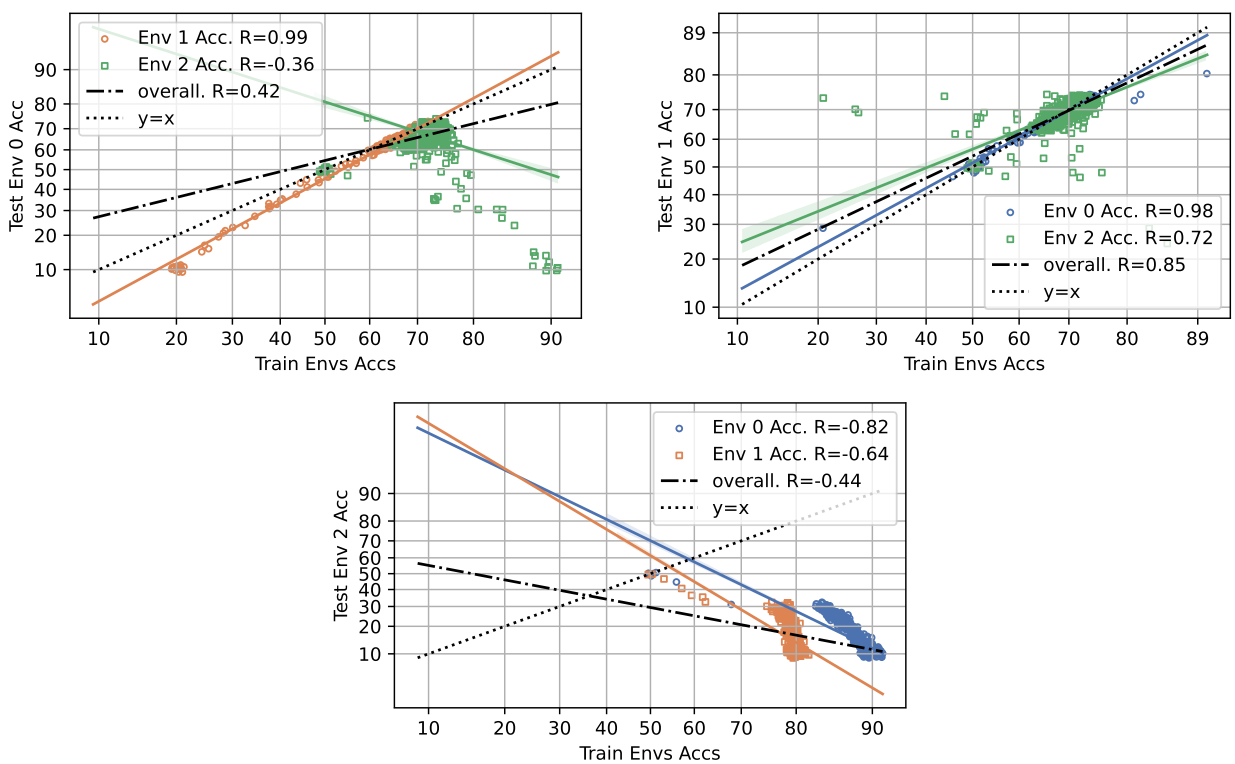


Figure 7: ColoredMNIST: Train Env Accuracy vs. Test Env Accuracy.

C.3 Spawrious

Spawrious (Lynch et al., 2023). The Spawrious image classification benchmark suite consists of six different datasets, including one-to-one (O2O) spurious correlations, where a single spurious attribute correlates with a binary label, and many-to-many (M2M) spurious correlations across multiple classes and spurious attributes. Each benchmark task is proposed with three difficulty levels: Easy, Medium, and Hard. The dataset contains images of four dog breeds $c \in \{bulldog, dachshund, labrador, corgi\}$ found in six backgrounds $b \in \{beach, desert, dirt, jungle, mountain, sand\}$. Images are generated using text-to-image models and filtered using an image-to-text model for quality control. This benchmark suite consists of 152,064 images of dimensions (3, 224, 224).

For the O2O task, the class (dog breed) and background combinations are sampled such that μ% of the images per class contain a spurious background b^{sp} and $(100 - \mu)$% contain a generic background b^{ge}. While the generic background is held constant for each class, each spurious background is observed in only one class ($p_{train}(b_i^{sp} \mid c_j) = 1$ if $i = j$ and 0 if $i \neq j$). Two separate training domains are defined by varying the value of μ. These induced spurious correlations are reverted to yield a test domain with an unseen class-background pair for each class ($p_{test}(b_i \mid c_i) = 1$).

For the M2M task, disjoint class and background groups are constructed $\mathcal{B}_1, \mathcal{B}_2, \mathcal{C}_1, \mathcal{C}_2$, each with two elements. To introduce the training domains, class-background combinations (c, b) are selected with $c \in \mathcal{C}_i$ and $b \in \mathcal{B}_i$. Each training domain consists of a single background per class such that $p^e_{train}(b_k \mid c_k) = e$, with domain index $e \in \{0, 1\}$, $b_k \in \mathcal{B}_i$, $c_k \in \mathcal{C}_i$. In contrast, the test domain is generated by selecting combinations from $c \in \mathcal{C}_i$ and $b \in \mathcal{B}_j$ with $i \neq j$ and sampling backgrounds such that $p_{test}(b_1 \mid c_k) = p_{test}(b_2 \mid c_k) = 0.5$ for $c_k \in \mathcal{C}_i$, $\{b_1, b_2\} = \mathcal{B}_j$.

The difficulty level (Easy, Medium, Hard) differs due to the splits in the available class-background combinations. These splits were empirically determined, and the full details of the final data combinations are found in Table 2 of Lynch et al. (2023).

Discussion. We observed that test environments 1 and 2 have a strong correlation between ID and OOD accuracy and have a slope of 1, making them misspecified for benchmarking domain generalization. Appropriately, Lynch et al. (2023) propose transferring to test environment '0' as the spurious correlation task. The correlation for test environment 0 is much weaker than the others, indicating that ID improvement does not as strongly imply OOD improvement. While there is still a positive linear correlation, the interpretation of these benchmarking results is informative because of the knowledge of the spurious correlation mechanism. Lynch et al. (2023) give examples of informative analysis of benchmarking results on this dataset. Notably, the O2O_easy setting has a weaker correlation by design, and the accuracy on the line strength increases. We see similar behavior for the M2M_ setting. However, this task is much harder than the O2O task, which is reflected in weaker accuracy on the line.

C.3.1 SpawriousO2O Easy

Table 6: SpawriousO2O_easy ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.48	-0.29	0.74	0.00	0.04
Env 1 acc	1.05	-0.13	0.98	0.00	0.02
Env 2 acc	0.95	-0.11	0.97	0.00	0.02

Table 7: SpawriousO2O_easy ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.50	-0.33	0.75	0.00	0.04
Env 0 acc	Env 2 acc	0.47	-0.23	0.72	0.00	0.04
Env 1 acc	Env 0 acc	1.09	-0.33	0.93	0.00	0.04
Env 1 acc	Env 2 acc	1.01	0.11	0.98	0.00	0.02
Env 2 acc	Env 0 acc	0.93	-0.12	0.94	0.00	0.03
Env 2 acc	Env 1 acc	0.94	-0.02	0.98	0.00	0.01

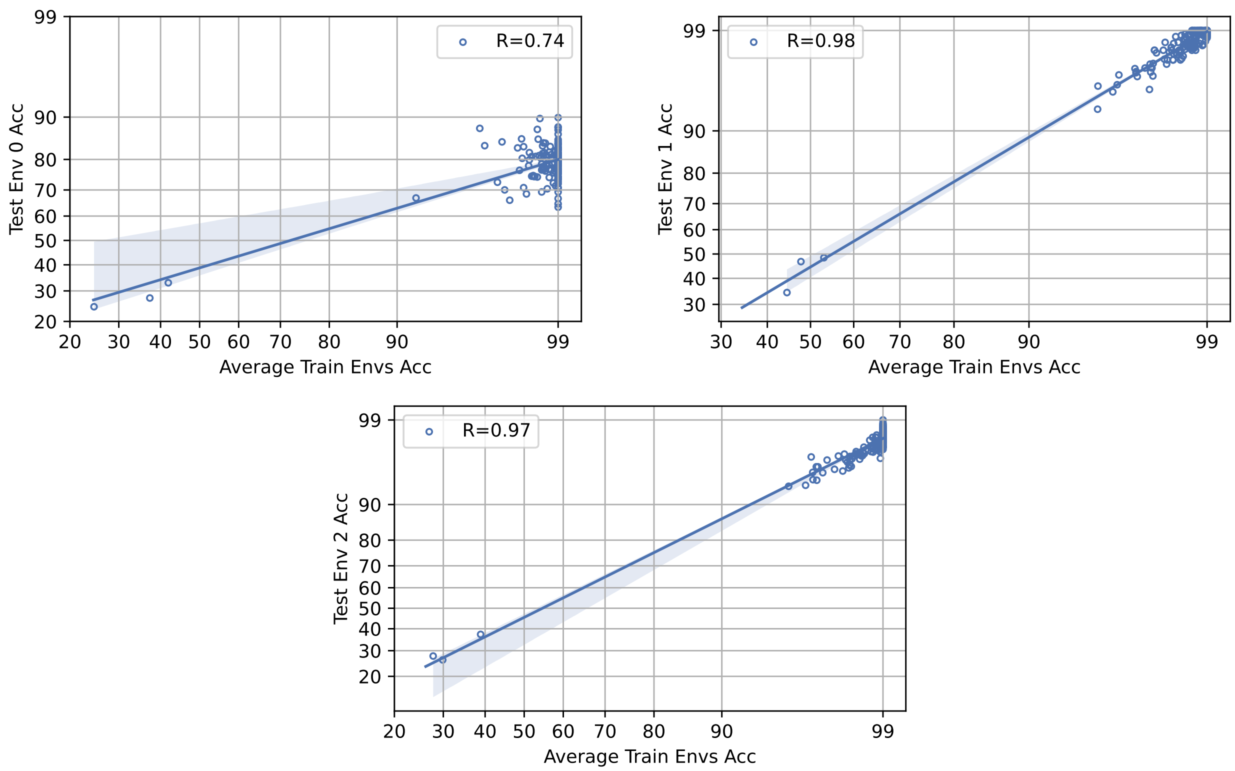


Figure 8: SpawriousO2O easy: Average train Env Accuracy vs. Test Env Accuracy.

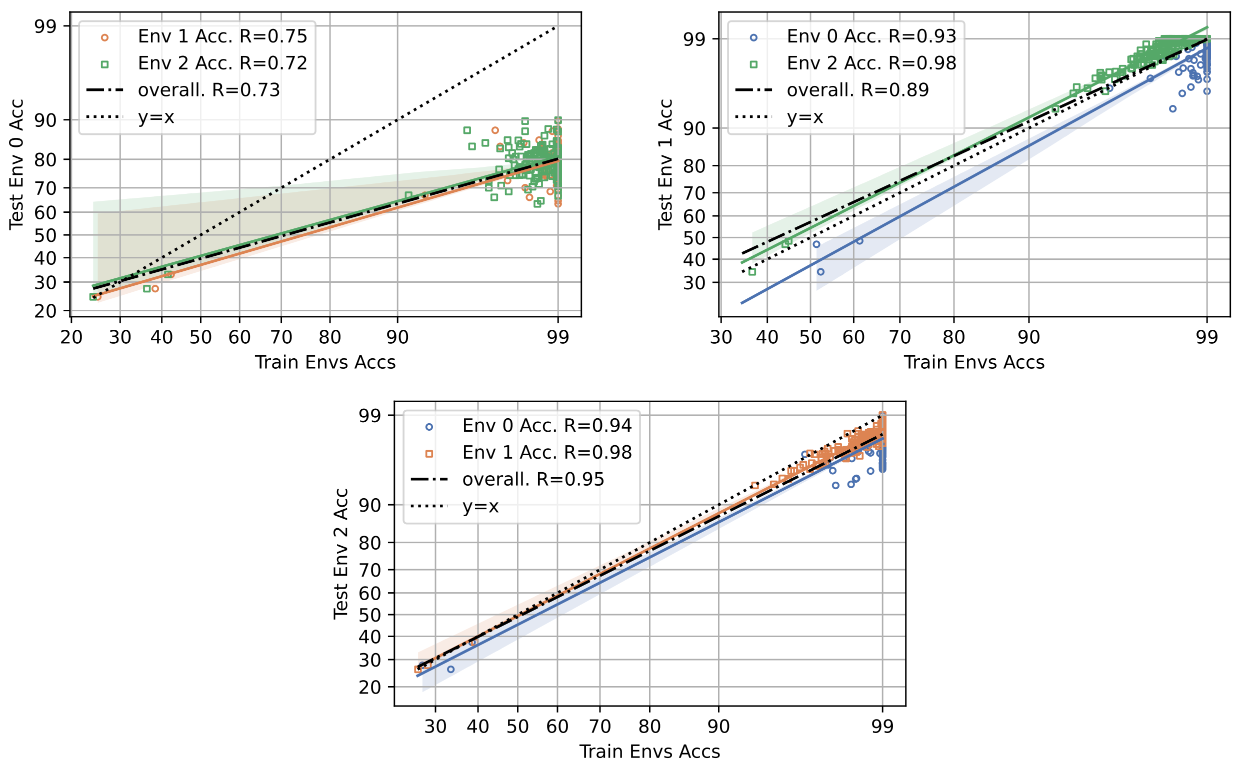


Figure 9: SpawriousO2O easy: Train Env Accuracy vs. Test Env Accuracy.

C.3.2 SpawriousO2O Hard

Table 8: SpawriousO2O_hard ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.32	-0.21	0.50	0.00	0.05
Env 1 acc	0.98	0.06	0.96	0.00	0.02
Env 2 acc	0.94	-0.07	0.96	0.00	0.02

Table 9: SpawriousO2O_hard ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.32	-0.23	0.49	0.00	0.05
Env 0 acc	Env 2 acc	0.32	-0.19	0.50	0.00	0.04
Env 1 acc	Env 0 acc	0.90	0.14	0.89	0.00	0.04
Env 1 acc	Env 2 acc	1.01	0.12	0.97	0.00	0.02
Env 2 acc	Env 0 acc	0.92	-0.05	0.93	0.00	0.03
Env 2 acc	Env 1 acc	0.92	0.01	0.97	0.00	0.02

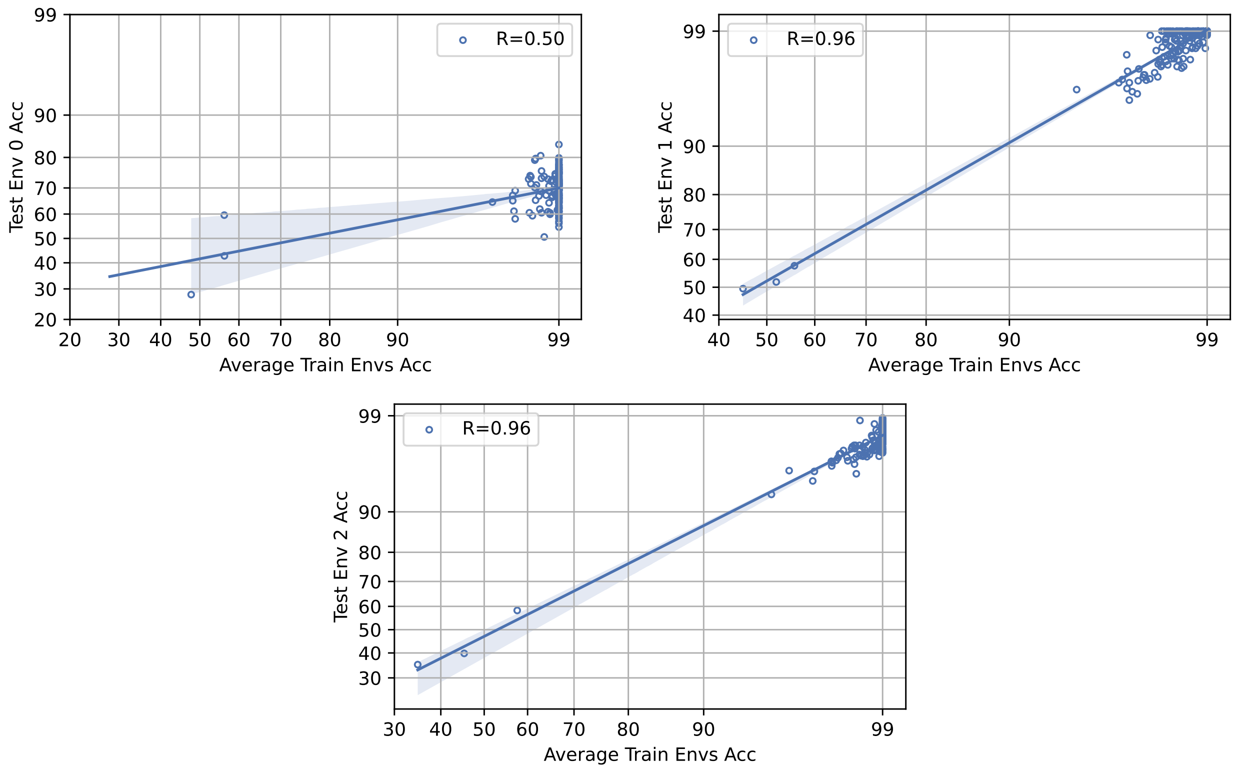


Figure 10: SpawriousO2O hard: Average train Env Accuracy vs. Test Env Accuracy.

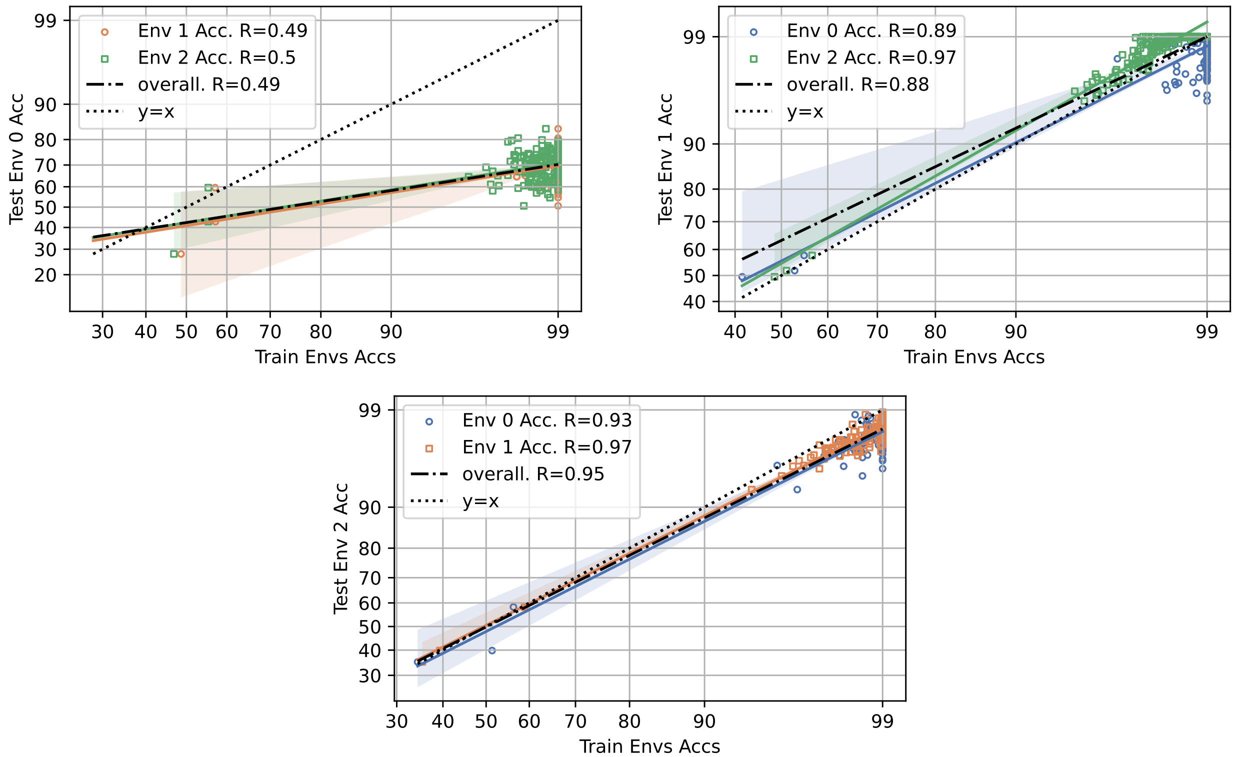


Figure 11: SpawriousO2O hard: Train Env Accuracy vs. Test Env Accuracy.

C.3.3 SpawriousM2M Easy

Table 10: SpawriousM2M_easy ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.34	0.26	0.60	0.00	0.01
Env 1 acc	0.65	-0.08	0.95	0.00	0.00
Env 2 acc	0.65	0.02	0.93	0.00	0.00

Table 11: SpawriousM2M_easy ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.35	0.23	0.61	0.00	0.01
Env 0 acc	Env 2 acc	0.32	0.29	0.58	0.00	0.01
Env 1 acc	Env 0 acc	0.64	-0.09	0.95	0.00	0.00
Env 1 acc	Env 2 acc	0.63	-0.05	0.94	0.00	0.00
Env 2 acc	Env 0 acc	0.67	-0.02	0.94	0.00	0.00
Env 2 acc	Env 1 acc	0.61	0.08	0.90	0.00	0.01

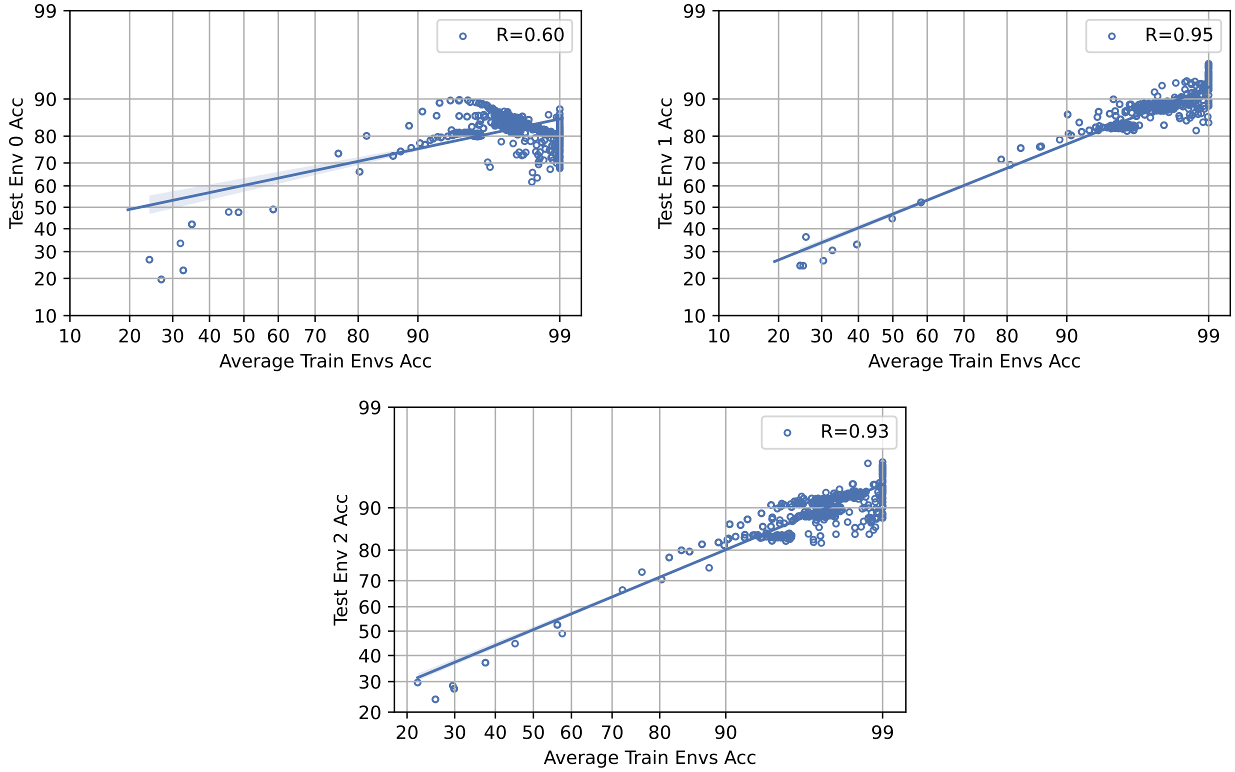


Figure 12: SpawriousO2O easy: Average train Env Accuracy vs. Test Env Accuracy.

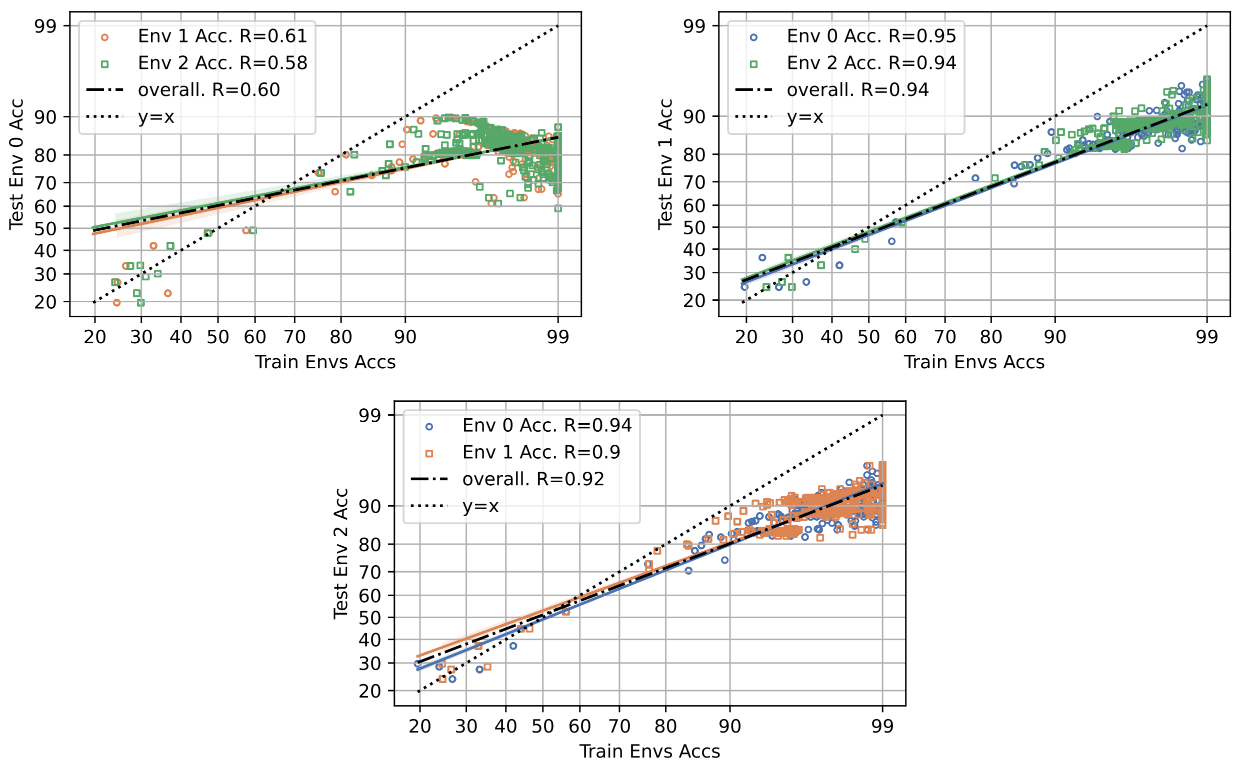


Figure 13: SpawriousO2O easy: Train Env Accuracy vs. Test Env Accuracy.

C.3.4 SpawriousM2M Hard

Table 12: SpawriousM2M_hard ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.16	-0.04	0.29	0.00	0.01
Env 1 acc	0.76	-0.26	0.94	0.00	0.01
Env 2 acc	0.66	-0.10	0.91	0.00	0.01

Table 13: SpawriousM2M_hard ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.17	-0.07	0.33	0.00	0.01
Env 0 acc	Env 2 acc	0.14	-0.02	0.27	0.00	0.01
Env 1 acc	Env 0 acc	0.78	-0.27	0.95	0.00	0.00
Env 1 acc	Env 2 acc	0.73	-0.23	0.92	0.00	0.01
Env 2 acc	Env 0 acc	0.68	-0.11	0.92	0.00	0.01
Env 2 acc	Env 1 acc	0.62	-0.08	0.89	0.00	0.01

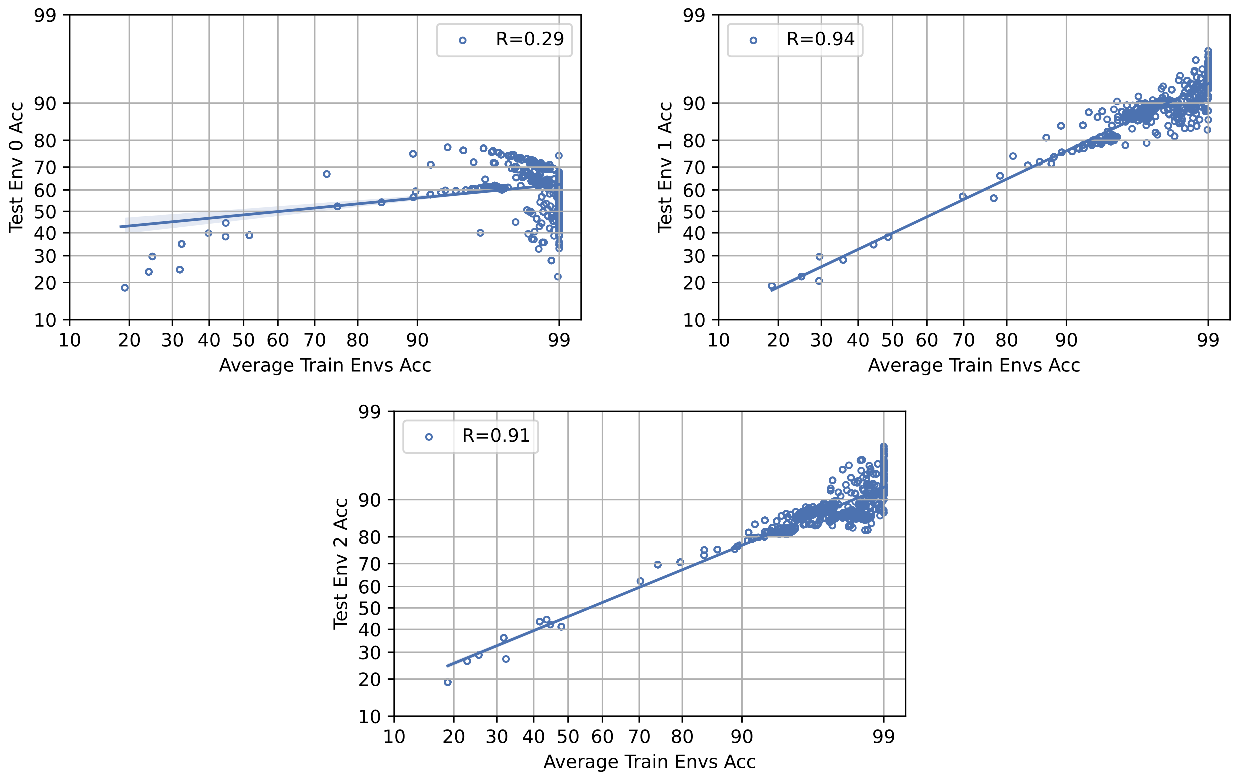


Figure 14: SpawriousM2M Hard: Average train Env Accuracy vs. Test Env Accuracy.

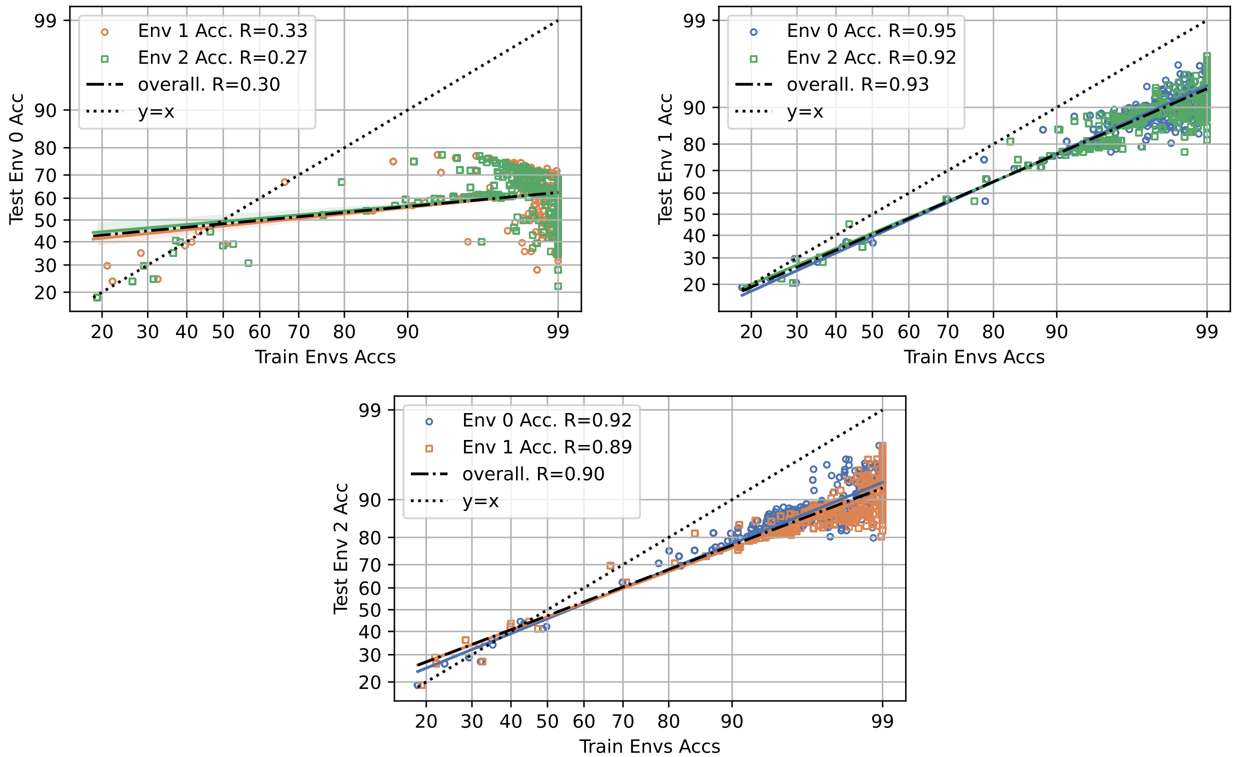


Figure 15: SpawriousM2M Hard: Train Env Accuracy vs. Test Env Accuracy.

C.4 PACS

PACS (Li et al., 2017). A dataset comprised of four domains $d \in \{art, cartoons, photos, sketches\}$. This dataset contains 9,991 examples of dimension (3, 224, 224) and 7 classes.

Discussion. In general, we find that PACS does not strongly represent worst-case shifts for any split. Our results suggest that this benchmark may not accurately benchmark an algorithm's ability to give models free of spurious correlations.

Table 14: PACS ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.74	-0.31	0.98	0.00	0.00
Env 1 acc	0.68	-0.68	0.84	0.00	0.01
Env 2 acc	1.00	0.32	0.86	0.00	0.01
Env 3 acc	0.76	-0.87	0.86	0.00	0.01

Table 15: PACS ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.71	-0.10	0.96	0.00	0.01
Env 0 acc	Env 2 acc	0.64	-0.47	0.91	0.00	0.01
Env 0 acc	Env 3 acc	0.64	0.14	0.90	0.00	0.01
Env 1 acc	Env 0 acc	0.75	-0.59	0.89	0.00	0.01
Env 1 acc	Env 2 acc	0.51	-0.67	0.71	0.00	0.01
Env 1 acc	Env 3 acc	0.71	-0.29	0.98	0.00	0.00
Env 2 acc	Env 0 acc	1.04	0.23	0.87	0.00	0.01
Env 2 acc	Env 1 acc	0.90	0.45	0.82	0.00	0.02
Env 2 acc	Env 3 acc	0.78	0.76	0.75	0.00	0.02
Env 3 acc	Env 0 acc	0.76	-0.79	0.83	0.00	0.01
Env 3 acc	Env 1 acc	0.80	-0.75	0.92	0.00	0.01
Env 3 acc	Env 2 acc	0.50	-0.83	0.64	0.00	0.02

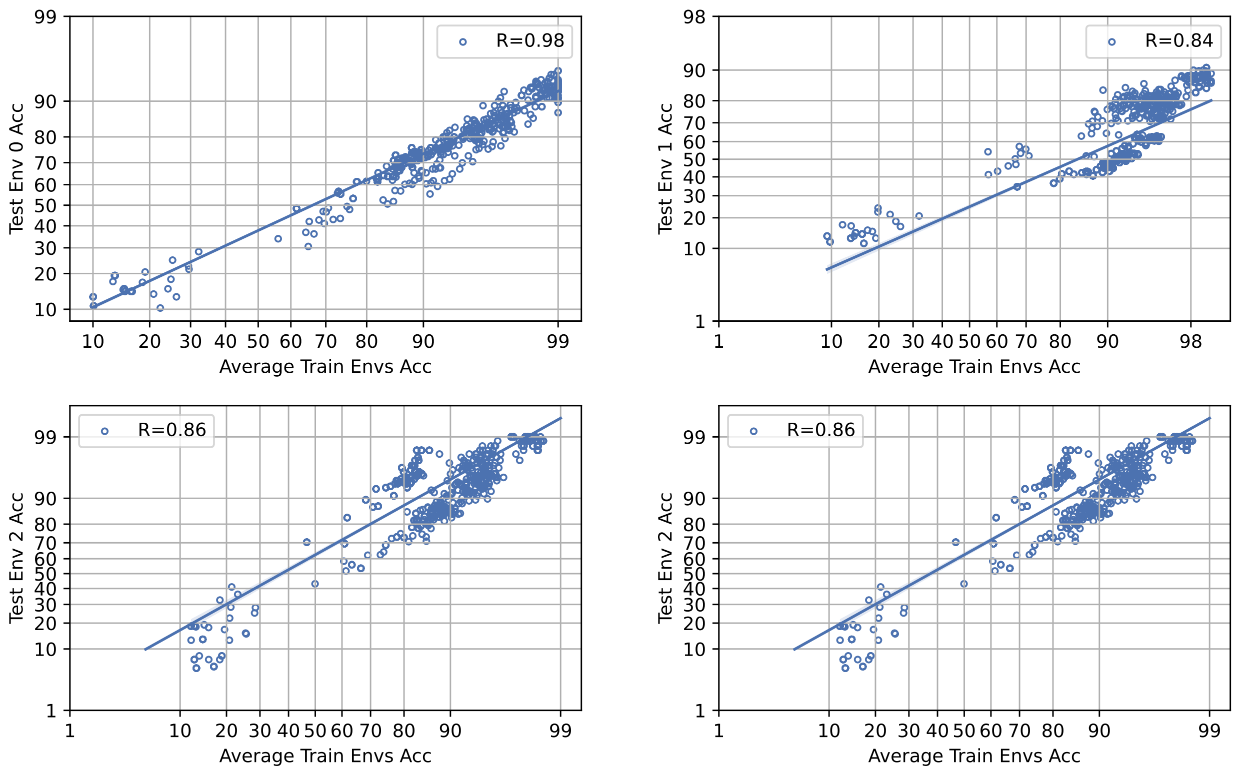


Figure 16: PACS: Average train Env Accuracy vs. Test Env Accuracy.

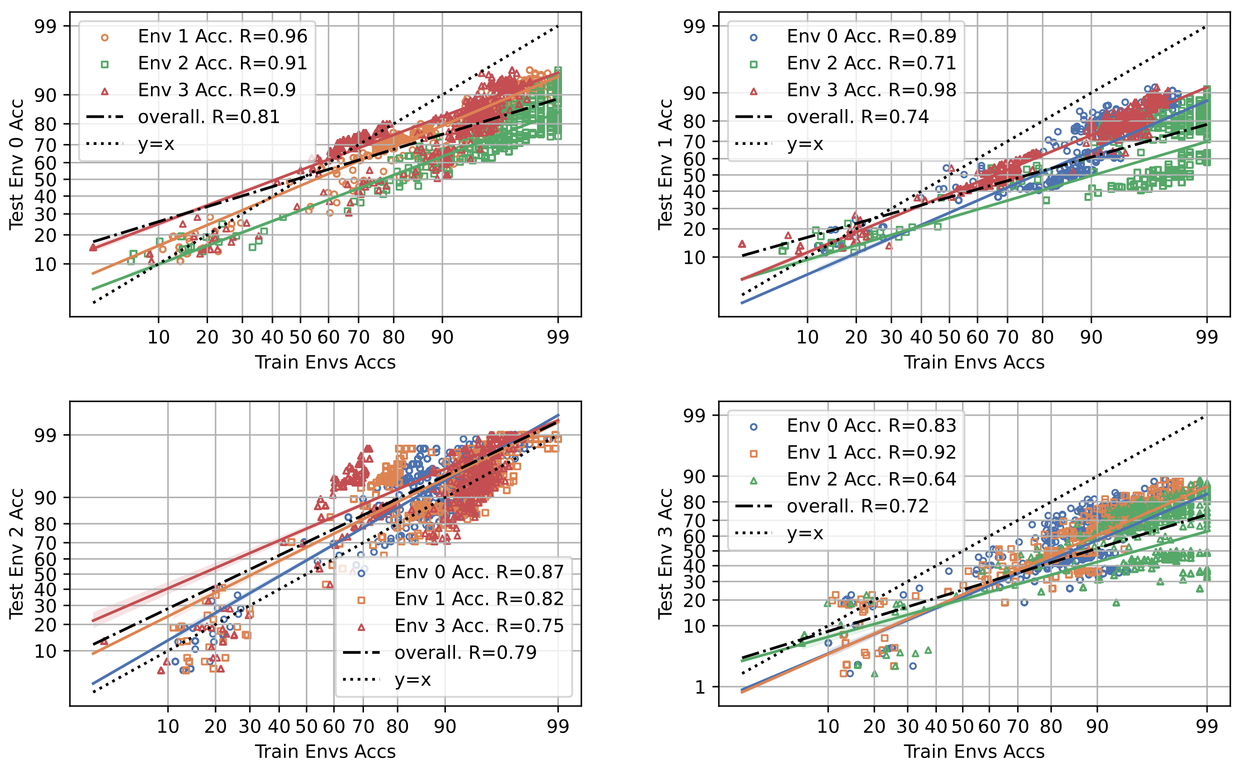


Figure 17: PACS: Train Env Accuracy vs. Test Env Accuracy.

C.5 Terra Incognita

Terra Incognita (Beery et al., 2018). A dataset that contains photographs of wild animals taken by camera traps at locations $d \in \{L100, L38, L43, L46\}$. This dataset contains 24,788 examples of dimensions (3, 224, 224) and 10 classes: Bird, Bobcat, Cat, Coyote, Dog, Empty, Fox, Horse, Mouse, Opossum, Rabbit, Raccoon, Rat, Skunk, Squirrel, Weasel.

Discussion. In general, we find that Terra Incognita does not strongly represent worst-case shifts for any split. Ahuja et al. (2021) consider Terra Incognita domain-general features to be fully informative, i.e., labels did not need to rely on spurious features such as the background to generate labels. Our results suggest that this benchmark may not accurately benchmark an algorithm's ability to give models free of spurious correlations. For Env 1, we observe that the slope of the line varies quite a bit. Particularly, there is a near-zero slope for models greater than 80% accuracy.

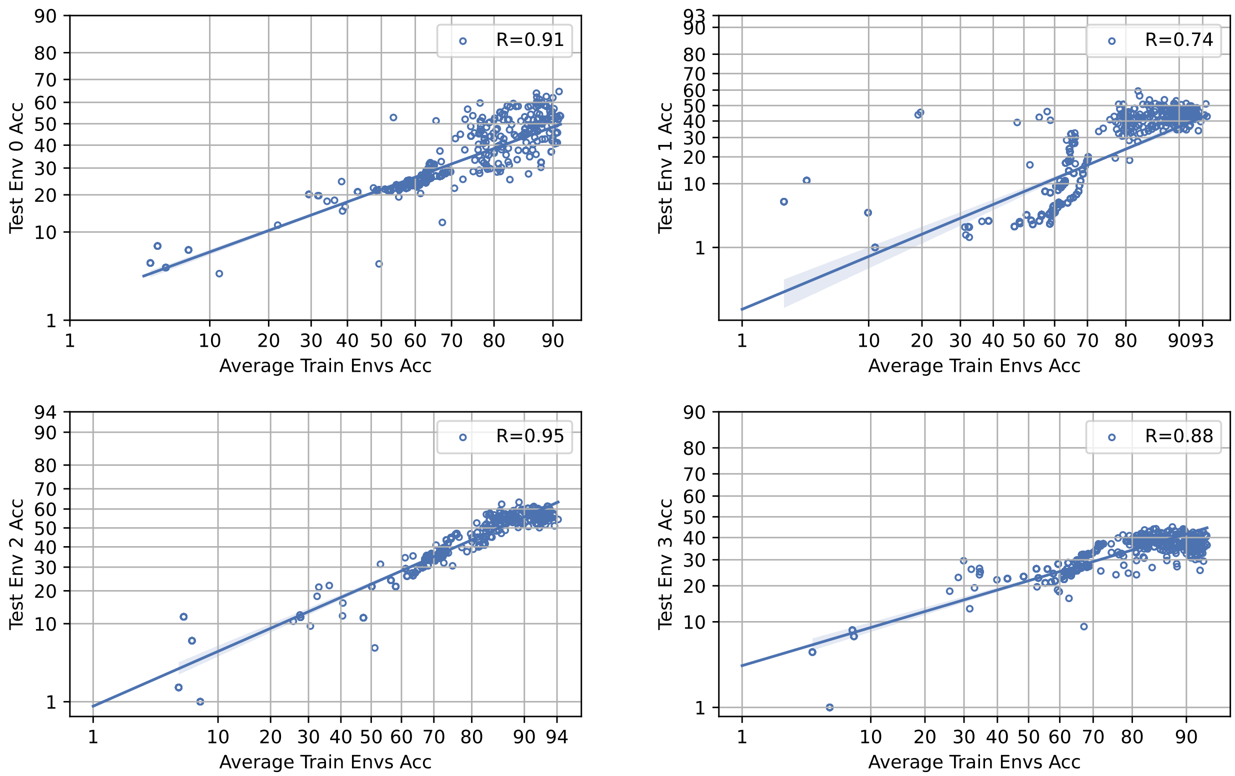


Figure 18: Terra Incognita: Average train Env Accuracy vs. Test Env Accuracy.

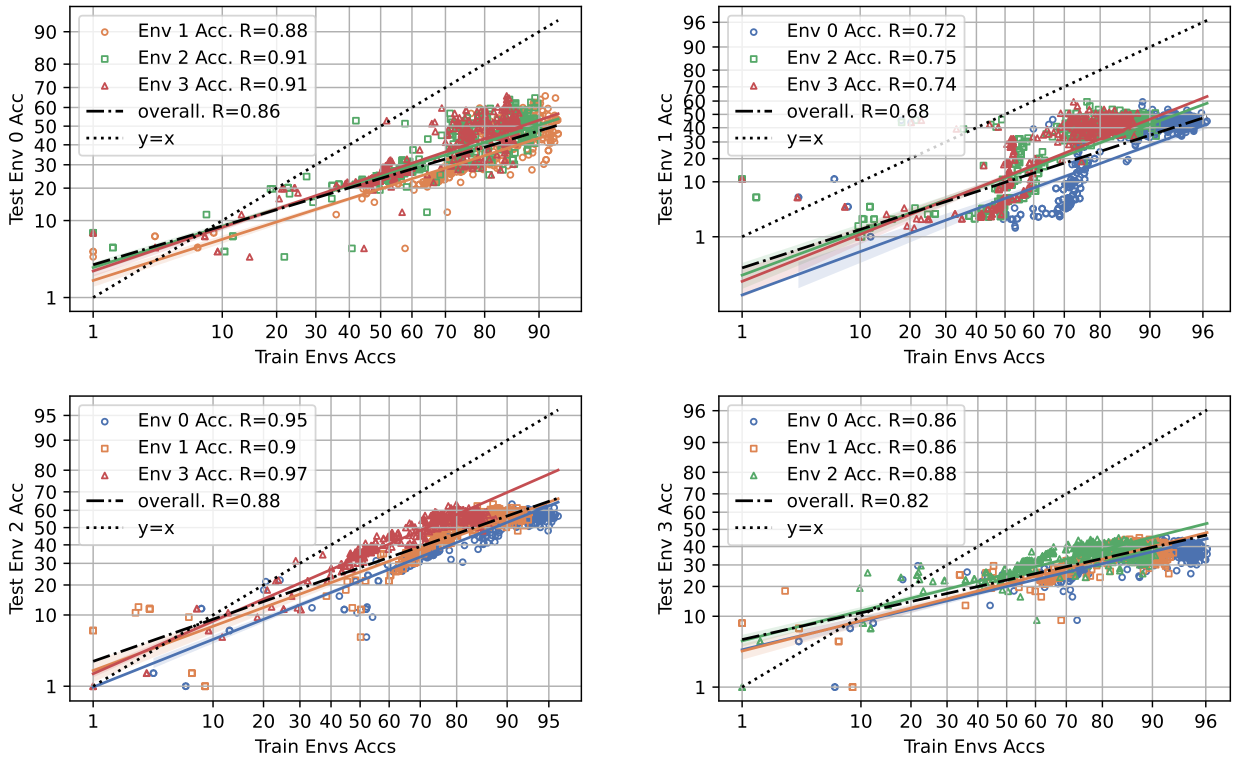


Figure 19: Terra Incognita: Train Env Accuracy vs. Test Env Accuracy.

C.6 Camelyon

Camelyon (Bandi et al., 2018; Koh et al., 2021). A dataset that contains histopathological images of lymph node tissue, collected from two hospitals, denoted as Hospital A, Hospital B. This dataset contains 327,680 examples of dimension (3, 96, 96) and 2 classes (tumor, non-tumor).

Discussion. We find that overall, there is a strong correlation between ID and OOD accuracy. However, we observe that for some ID/OOD splits, a regime of training accuracy has a negative correlation (environments 0 and 2), suggesting that within a certain accuracy range, these splits may be well-specified for benchmarking spurious correlations for models in the regime with negative correlation. This highlights the importance of qualitative evaluation as opposed to quantitative evaluation.

Table 16: WILDSCamelyon ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.78	0.33	0.90	0.00	0.01
Env 1 acc	0.71	-0.00	0.88	0.00	0.01
Env 2 acc	0.62	0.49	0.78	0.00	0.01
Env 3 acc	0.63	0.49	0.88	0.00	0.01
Env 4 acc	0.63	0.40	0.78	0.00	0.01

Table 17: WILDSCamelyon ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.73	0.44	0.88	0.00	0.01
Env 0 acc	Env 2 acc	0.79	0.25	0.90	0.00	0.01
Env 0 acc	Env 3 acc	0.83	0.17	0.90	0.00	0.01
Env 0 acc	Env 4 acc	0.74	0.28	0.91	0.00	0.01
Env 1 acc	Env 0 acc	0.71	-0.00	0.89	0.00	0.01
Env 1 acc	Env 2 acc	0.69	0.02	0.85	0.00	0.01
Env 1 acc	Env 3 acc	0.74	-0.05	0.89	0.00	0.01
Env 1 acc	Env 4 acc	0.69	-0.03	0.89	0.00	0.01
Env 2 acc	Env 0 acc	0.64	0.41	0.81	0.00	0.01
Env 2 acc	Env 1 acc	0.59	0.58	0.74	0.00	0.01
Env 2 acc	Env 3 acc	0.67	0.37	0.79	0.00	0.01
Env 2 acc	Env 4 acc	0.61	0.44	0.81	0.00	0.01
Env 3 acc	Env 0 acc	0.66	0.41	0.90	0.00	0.01
Env 3 acc	Env 1 acc	0.57	0.64	0.84	0.00	0.01
Env 3 acc	Env 2 acc	0.63	0.47	0.87	0.00	0.01
Env 3 acc	Env 4 acc	0.61	0.45	0.88	0.00	0.01
Env 4 acc	Env 0 acc	0.63	0.35	0.79	0.00	0.01
Env 4 acc	Env 1 acc	0.60	0.49	0.74	0.00	0.01
Env 4 acc	Env 2 acc	0.62	0.39	0.79	0.00	0.01
Env 4 acc	Env 3 acc	0.67	0.29	0.78	0.00	0.01

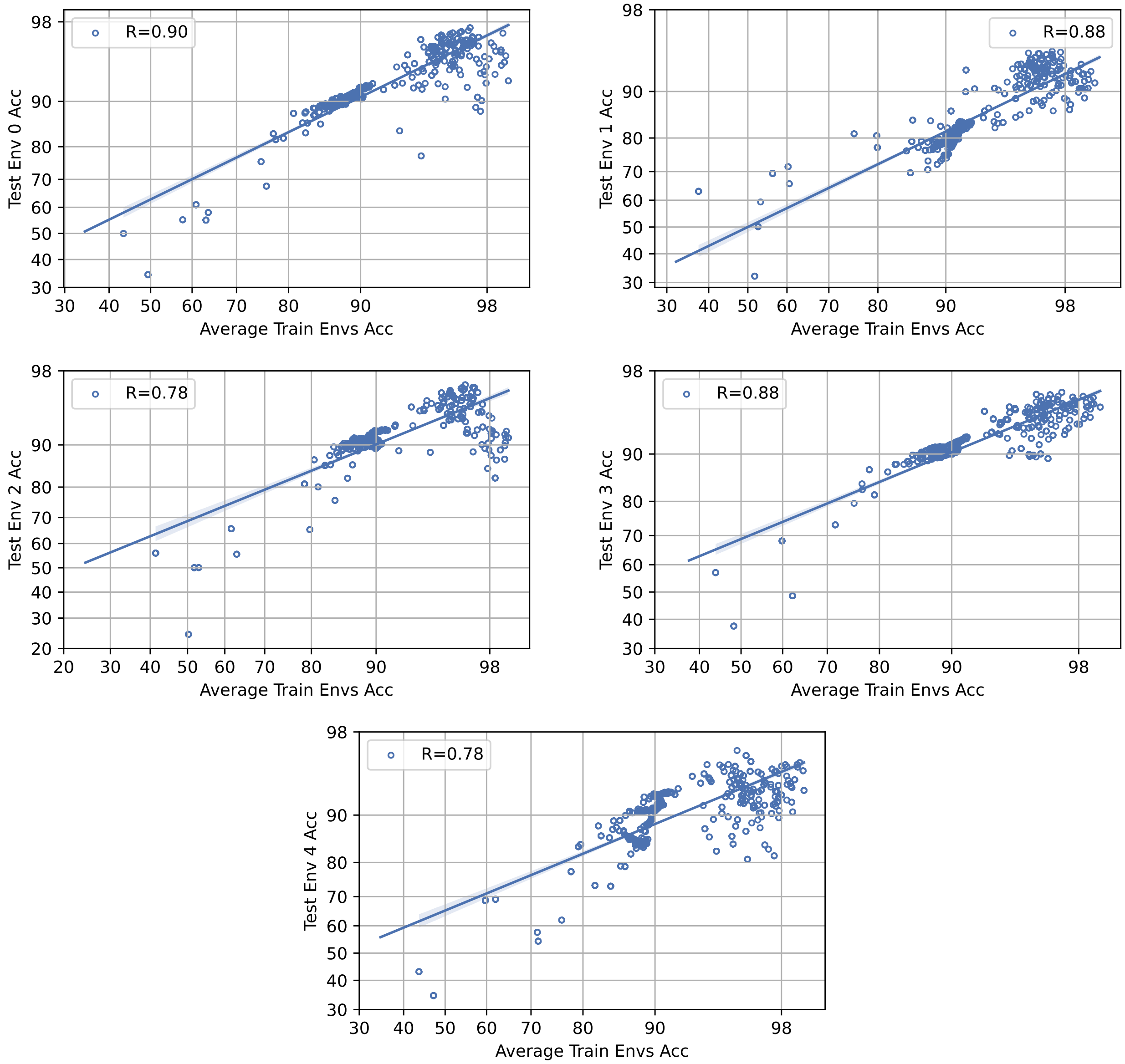


Figure 20: Camelyon: Average train Env Accuracy vs. Test Env Accuracy.

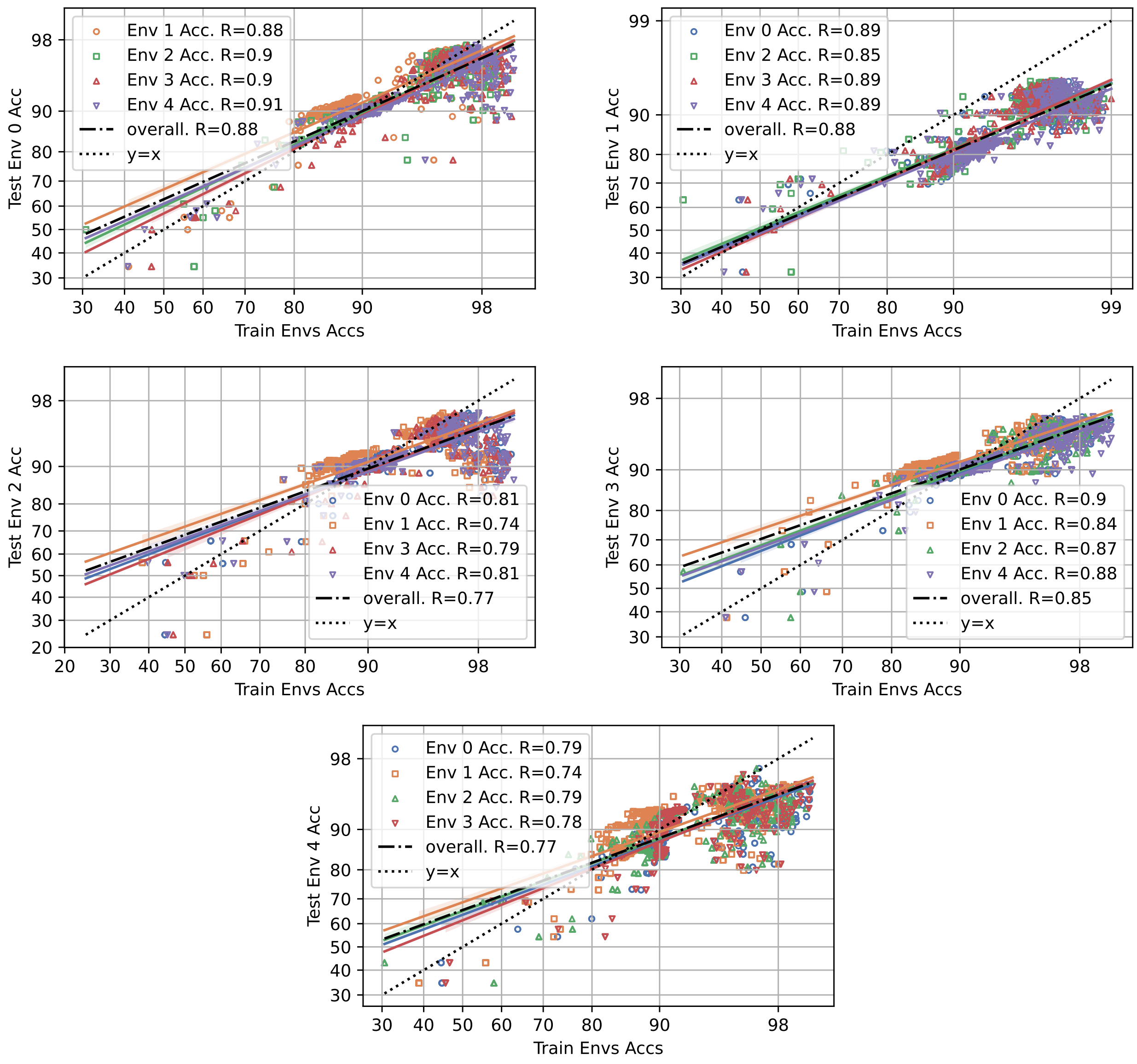


Figure 21: Camelyon: Train Env Accuracy vs. Test Env Accuracy.

C.7 Covid-CXR

Covid-CXR (Alzate-Grisales et al., 2022; Cohen et al., 2020b; Tabik et al., 2020; Tahir et al., 2021; Suwalska et al., 2023). A dataset that aggregates five different benchmark Covid-19 datasets, represented as the domain $d \in$ {Cov-Caldas (Columbia), Covid-Chest-X-Ray (Global), Covid-GR (Spain), Covid-Qu-Ex (Global), and PolCovid (Poland)}. These open-source Covid-19 datasets are collected from around the globe, accessible online. We are provided with *chest X-ray image (x)* and *binary diagnosis (y)*. The objective is to maintain consistent performance across different datasets d, which may incorporate data from singular or multiple sources.

Dataset	# Train	# Test	% Pos Train	% Neg Train	% Pos Test	% Neg Test
Cov-Caldas	3247	688	0.566	0.434	0.565	0.435
Covid-Chest-X-Ray	625	157	0.376	0.624	0.376	0.624
Covid-GR	681	171	0.501	0.499	0.497	0.503
Covid-Qu-Ex	21715	6788	0.353	0.647	0.353	0.647
PolCovid	4343	450	0.249	0.751	0.333	0.667

Table 18: Composition for Covid-19 Datasets.

- Cov-Caldas (Columbia, Alzate-Grisales et al. (2022)): This dataset was sourced from a single institution, S.E.S. Hospital Universitario de Caldas, located in the State of Caldas, Colombia. Labels were assigned based on positive results from conventional laboratory tests, such as PCR.
- Covid-Chest-X-Ray (Global, Cohen et al. (2020b)): This dataset, compiled from web sources, publications, and volunteer contributions, includes data on five types of pneumonia and Covid-19, along with metadata such as sex, age, and symptoms. The images are sourced from medical websites and are part of an open-source public project, where contributors can submit pull requests to add new images. The data is compiled from 138 unique locations.
- Covid-GR (Spain, Tabik et al. (2020)): In collaboration with four expert radiologists from Hospital Universitario Clínico San Cecilio in Granada, Spain, the authors developed a protocol for selecting and annotating chest X-ray (CXR) images for the dataset. A CXR image is labeled as Covid-19 positive if both the RT-PCR test and the radiologist's assessment confirm the diagnosis within 24 hours.
- Covid-Qu-Ex (Global, Tahir et al. (2021)): This dataset aggregates chest X-rays from six subdatasets, including the Covid-19 CXR dataset, RSNA CXR dataset (non-Covid infections and normal CXRs), Chest-Xray-Pneumonia dataset, PadChest dataset, Montgomery and Shenzhen CXR lung mask datasets, and QaTa-Cov19 CXR infection mask dataset. Designed to serve as a benchmark, it combines multiple publicly available datasets and repositories, which were originally dispersed and formatted differently. The authors performed quality control to ensure consistency. This dataset, specifically portions of their cited Covid-19 CXR dataset, overlaps with Covid-Chest-X-Ray (Env 1), thus deviating slightly from our standard experimental procedure of leave-one-domain-out.
- PolCovid (Poland, Suwalska et al. (2023)): Chest X-rays were collected from 15 Polish hospitals using a variety of devices and parameters due to differences in equipment between medical centers. The dataset includes patients with Covid-19, pneumonia, and healthy controls.

Experimental Details. Using the DomainBed suite by Gulrajani & Lopez-Paz (2020b), we employ the ResNet-50 architecture (with and without AugMix data augmentation), along with ResNet-18, DenseNet-121, and ConvNeXt-Tiny, on the Covid-CXR dataset. Each of the five datasets is treated as a distinct domain, d. For each model, we apply two pretrained variants using ImageNet: (i) Fine-tuned and (ii) Transfer learning.

Discussion. Covid-CXR is a real-world medical dataset containing chest X-ray images of Covid-19 patients, where spurious correlations emerge naturally due to the complex and multifaceted nature of medical data. These correlations are not explicitly observed or intentionally introduced but arise from underlying factors such as imaging artifacts, patient demographics, or co-occurring medical conditions that are often unmeasured or unaccounted for in the dataset. As observed, there exist strong, weak, and inverse correlations between ID and OOD performance. While environments 0, 2, and 4 are primarily positively correlated in OOD performance, environments 1 and 3 demonstrate the inverse accuracy-on-the-line phenomenon. Additionally, several observations show weak correlations, with slopes close to zero. Existing literature suggests that a horizontal line (i.e. slope of 0) is indicative of a severe distribution shift, where it prevents any meaningful transfer learning between the training data and the OOD data Teney et al. (2024). This could be attributed to the significantly more severe distribution shifts present in this dataset. These shifts, as discussed by Cohen et al. (2020a), may arise due to errors in labeling, discrepancies between institutions and radiologists, biases in clinical practices, and interobserver variability.

Moreover, Covid-Qu-Ex (Env 3), the largest and most diverse dataset of those explored in this work, demonstrates low OOD transfer accuracy. For OOD performance for ID Env 3, we observe slopes closest to 0, indicating a weak or near-zero correlation. The high ID accuracy (up to $\approx 99\%$) suggests that the model may be learning misleading features in OOD domains. However, for environments evaluated on this dataset, a strongly negative relationship is present, suggesting that improved ID performance may be associated with reduced OOD accuracy. In this dataset, the authors scale by concatenating existing chest X-Ray datasets. But, as discussed in Cohen et al. (2020a), Shen et al. (2024), and Teney et al. (2024), simply increasing the amount of data may not address the core issue of distribution shift, as more data could exacerbate overfitting to domain-specific artifacts or noise, rather than improving generalization. Shen et al. (2024) coins this as the *Data Addition Dilemma*, where adding data can both improve and worsen performance.

Table 19: CXR ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.54	-0.27	0.55	0.00	0.02
Env 1 acc	-0.38	0.13	-0.50	0.00	0.02
Env 2 acc	0.44	0.05	0.54	0.00	0.02
Env 3 acc	-0.60	0.56	-0.48	0.00	0.03
Env 4 acc	0.53	-0.04	0.31	0.00	0.04

Table 20: CXR ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.56	-0.23	0.57	0.00	0.02
Env 0 acc	Env 2 acc	0.31	-0.21	0.41	0.00	0.02
Env 0 acc	Env 3 acc	0.16	-0.26	0.84	0.00	0.00
Env 0 acc	Env 4 acc	0.43	-0.39	0.68	0.00	0.01
Env 1 acc	Env 0 acc	-0.43	0.09	-0.61	0.00	0.01
Env 1 acc	Env 2 acc	-0.16	0.06	-0.27	0.00	0.01
Env 1 acc	Env 3 acc	-0.07	0.06	-0.51	0.00	0.00
Env 1 acc	Env 4 acc	-0.21	0.12	-0.46	0.00	0.01
Env 2 acc	Env 0 acc	0.39	0.07	0.54	0.00	0.01
Env 2 acc	Env 1 acc	0.36	0.06	0.46	0.00	0.02
Env 2 acc	Env 3 acc	0.09	0.06	0.61	0.00	0.00
Env 2 acc	Env 4 acc	0.28	-0.04	0.51	0.00	0.01
Env 3 acc	Env 0 acc	-0.68	0.54	-0.49	0.00	0.03
Env 3 acc	Env 1 acc	-0.47	0.51	-0.46	0.00	0.02
Env 3 acc	Env 2 acc	-0.38	0.51	-0.37	0.00	0.02
Env 3 acc	Env 4 acc	-0.26	0.54	-0.29	0.00	0.02
Env 4 acc	Env 0 acc	-0.03	0.19	-0.02	0.48	0.04
Env 4 acc	Env 1 acc	0.56	-0.01	0.40	0.00	0.03
Env 4 acc	Env 2 acc	0.47	-0.07	0.37	0.00	0.03
Env 4 acc	Env 3 acc	0.10	0.04	0.36	0.00	0.01

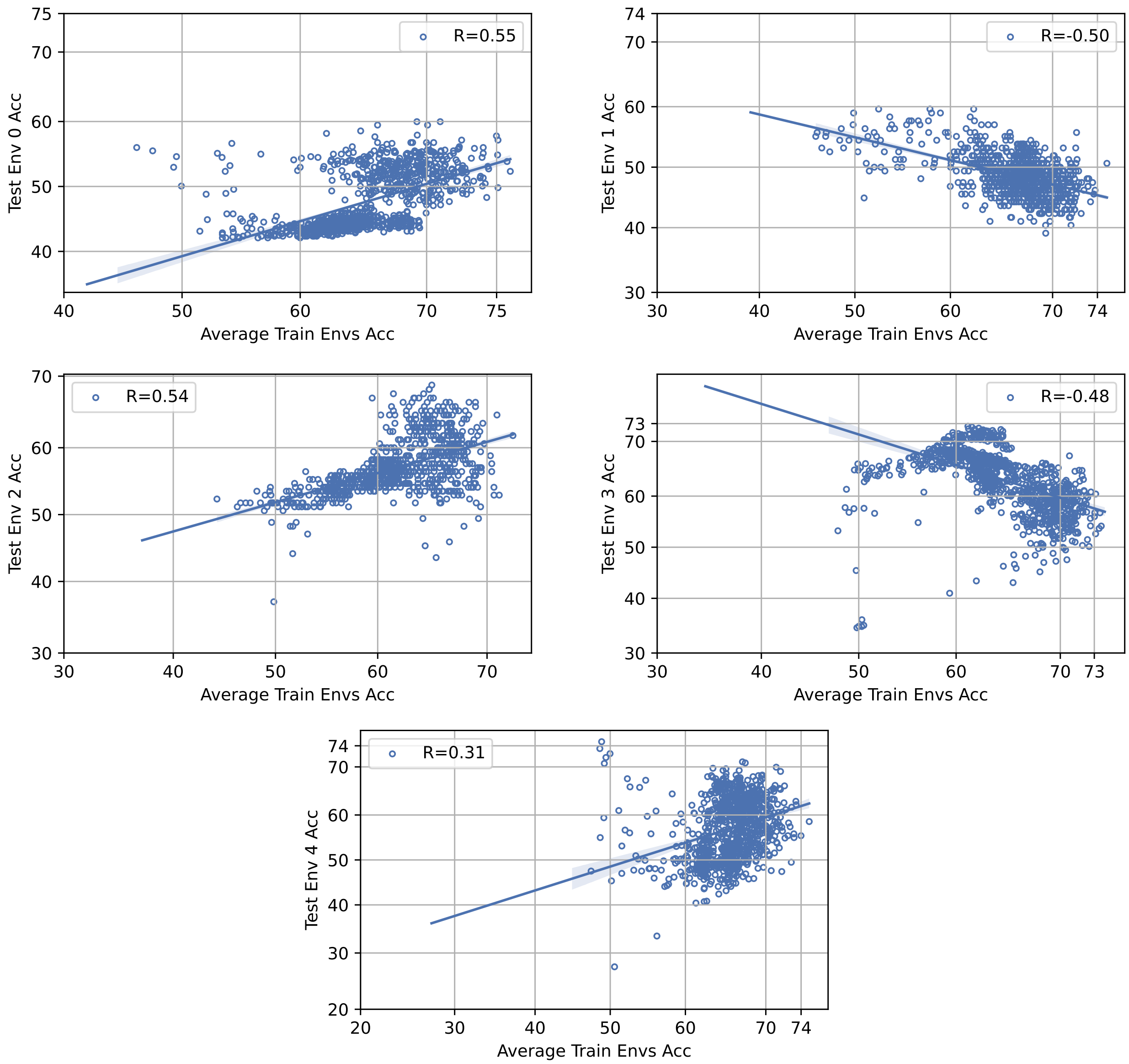


Figure 22: Covid-CXR: Average train Env Accuracy vs. Test Env Accuracy.

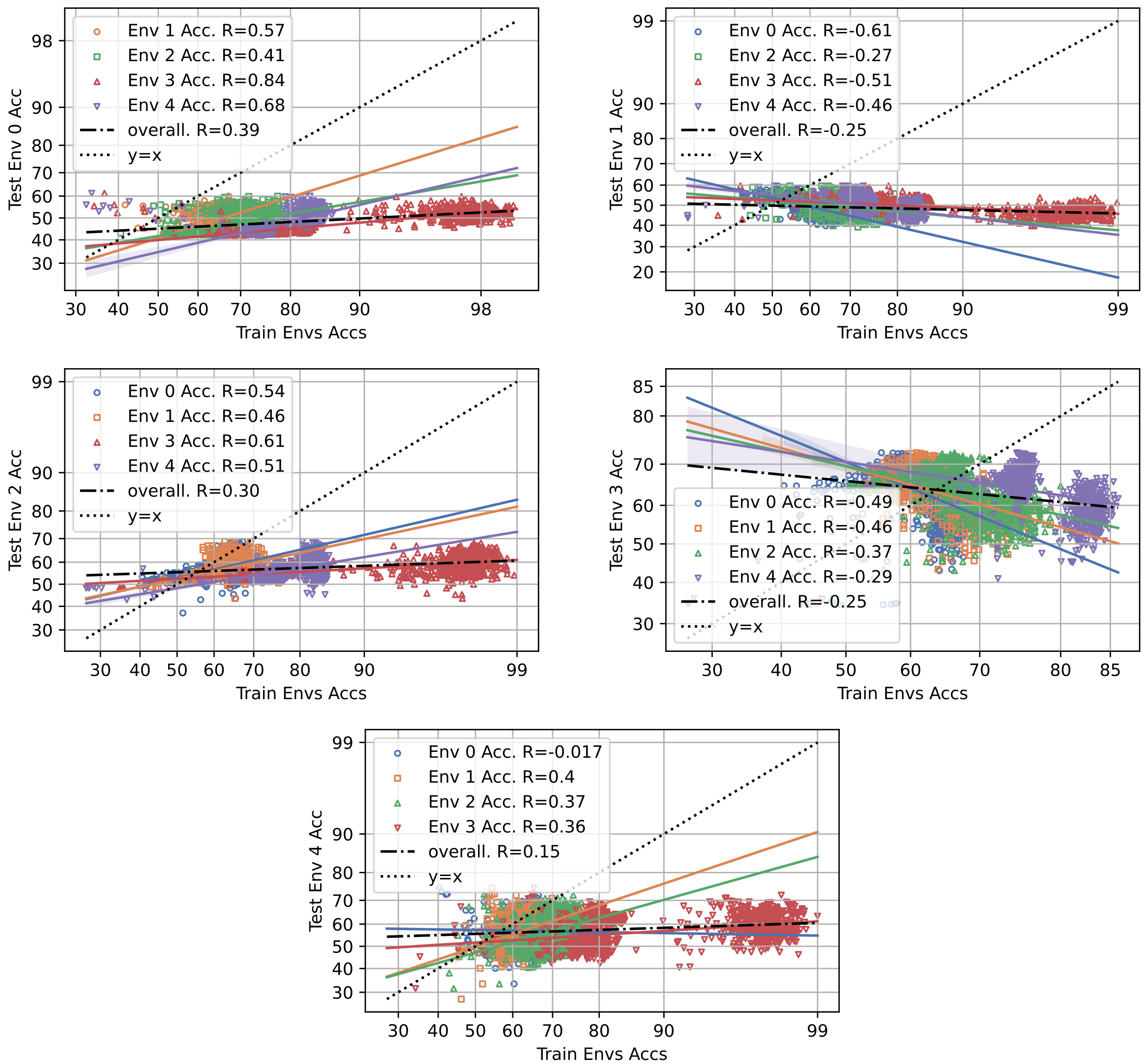


Figure 23: Covid-CXR: Train Env Accuracy vs. Test Env Accuracy.

C.8 FMoW

FMoW (Bandi et al., 2018; Koh et al., 2021). A dataset consisting of 141,696 RGB satellite images from 2022 - 2017 (resized to 224 x 224 pixels), where the label is one of 62 building or land use categories. This dataset simultaneously considers a domain generalization task, where two domains are defined by the year of image acquisition $t \in \{before2016, after2016\}$, and a subpopulation shift task, where the domains are denoted by the geographic region of the image $r \in \{Africa, Americas, Oceana, Asia, Europe\}$.

Discussion. We find that WILDFMoW has accuracy on the line for all splits, suggesting that this benchmark may be misspecified for benchmarking spurious correlations.

Table 21: WILDSFMoW ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.75	-0.62	0.98	0.00	0.00
Env 1 acc	0.78	-0.34	0.96	0.00	0.00
Env 2 acc	0.65	-0.48	0.94	0.00	0.00
Env 3 acc	0.83	-0.34	0.99	0.00	0.00
Env 4 acc	0.96	-0.18	0.99	0.00	0.00
Env 5 acc	0.76	-0.61	0.87	0.00	0.01

Table 22: WILDSFMoW ID vs. OOD properties.

OOD	ID	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.78	-0.52	0.98	0.00	0.00
Env 0 acc	Env 2 acc	0.69	-0.72	0.97	0.00	0.00
Env 0 acc	Env 3 acc	0.71	-0.52	0.97	0.00	0.00
Env 0 acc	Env 4 acc	0.48	-0.89	0.96	0.00	0.00
Env 0 acc	Env 5 acc	0.43	-1.11	0.94	0.00	0.00
Env 1 acc	Env 0 acc	0.75	-0.26	0.97	0.00	0.00
Env 1 acc	Env 2 acc	0.78	-0.43	0.96	0.00	0.00
Env 1 acc	Env 3 acc	0.80	-0.20	0.98	0.00	0.00
Env 1 acc	Env 4 acc	0.53	-0.62	0.95	0.00	0.00
Env 1 acc	Env 5 acc	0.49	-0.88	0.94	0.00	0.00
Env 2 acc	Env 0 acc	0.58	-0.50	0.94	0.00	0.00
Env 2 acc	Env 1 acc	0.70	-0.47	0.93	0.00	0.00
Env 2 acc	Env 3 acc	0.64	-0.49	0.92	0.00	0.00
Env 2 acc	Env 4 acc	0.43	-0.80	0.91	0.00	0.00
Env 2 acc	Env 5 acc	0.39	-1.01	0.92	0.00	0.00
Env 3 acc	Env 0 acc	0.74	-0.30	0.98	0.00	0.00
Env 3 acc	Env 1 acc	0.91	-0.25	0.99	0.00	0.00
Env 3 acc	Env 2 acc	0.79	-0.49	0.97	0.00	0.00
Env 3 acc	Env 4 acc	0.53	-0.67	0.96	0.00	0.00
Env 3 acc	Env 5 acc	0.49	-0.96	0.92	0.00	0.00
Env 4 acc	Env 0 acc	0.89	-0.13	0.98	0.00	0.00
Env 4 acc	Env 1 acc	1.02	-0.08	0.99	0.00	0.00
Env 4 acc	Env 2 acc	0.92	-0.33	0.98	0.00	0.00
Env 4 acc	Env 3 acc	0.95	-0.08	0.99	0.00	0.00
Env 4 acc	Env 5 acc	0.54	-0.87	0.90	0.00	0.00
Env 5 acc	Env 0 acc	0.74	-0.56	0.89	0.00	0.01
Env 5 acc	Env 1 acc	0.82	-0.54	0.85	0.00	0.01
Env 5 acc	Env 2 acc	0.70	-0.74	0.84	0.00	0.01
Env 5 acc	Env 3 acc	0.74	-0.55	0.85	0.00	0.01
Env 5 acc	Env 4 acc	0.52	-0.90	0.86	0.00	0.00

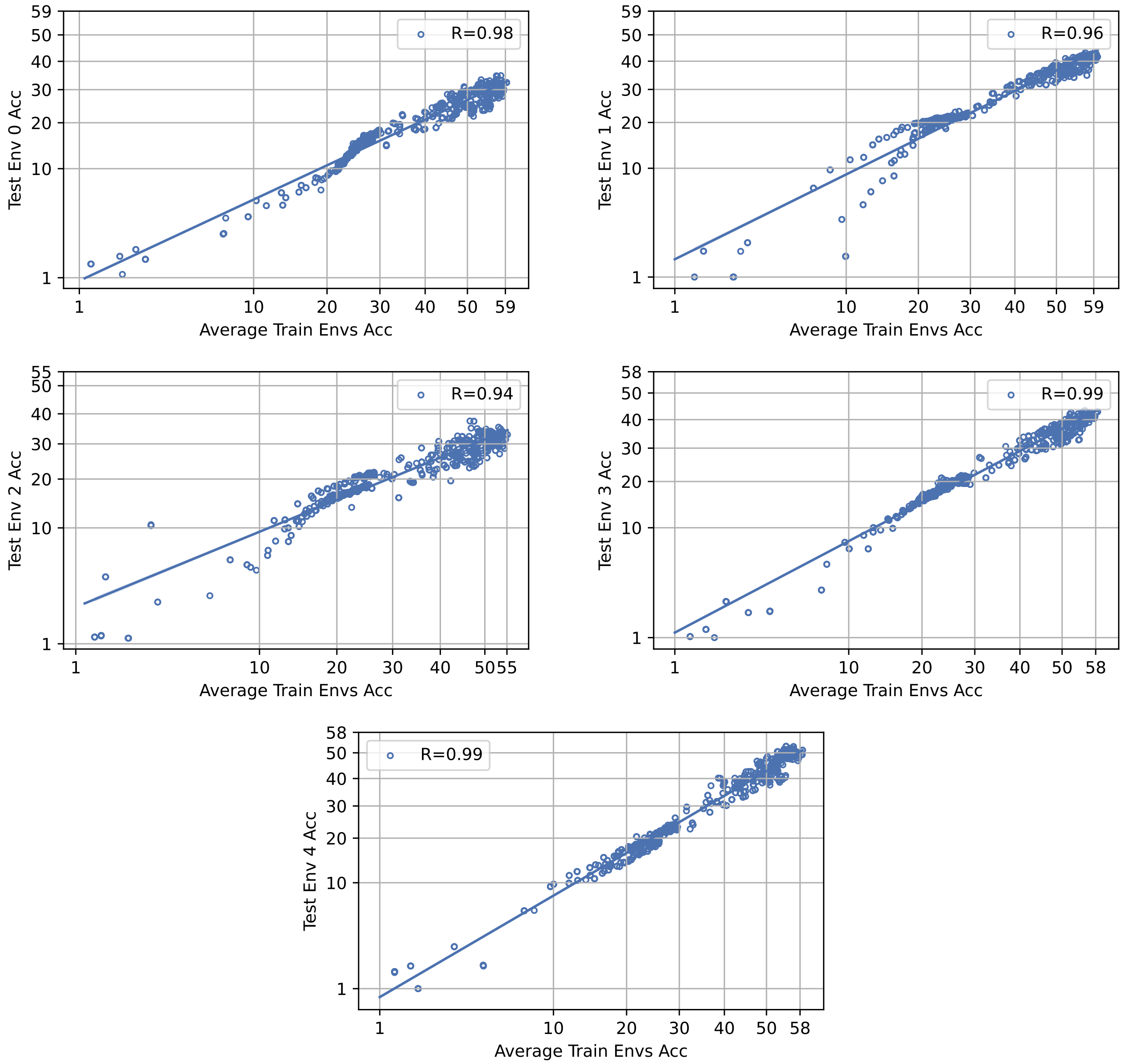


Figure 24: FMoW: Average train Env Accuracy vs. Test Env Accuracy.

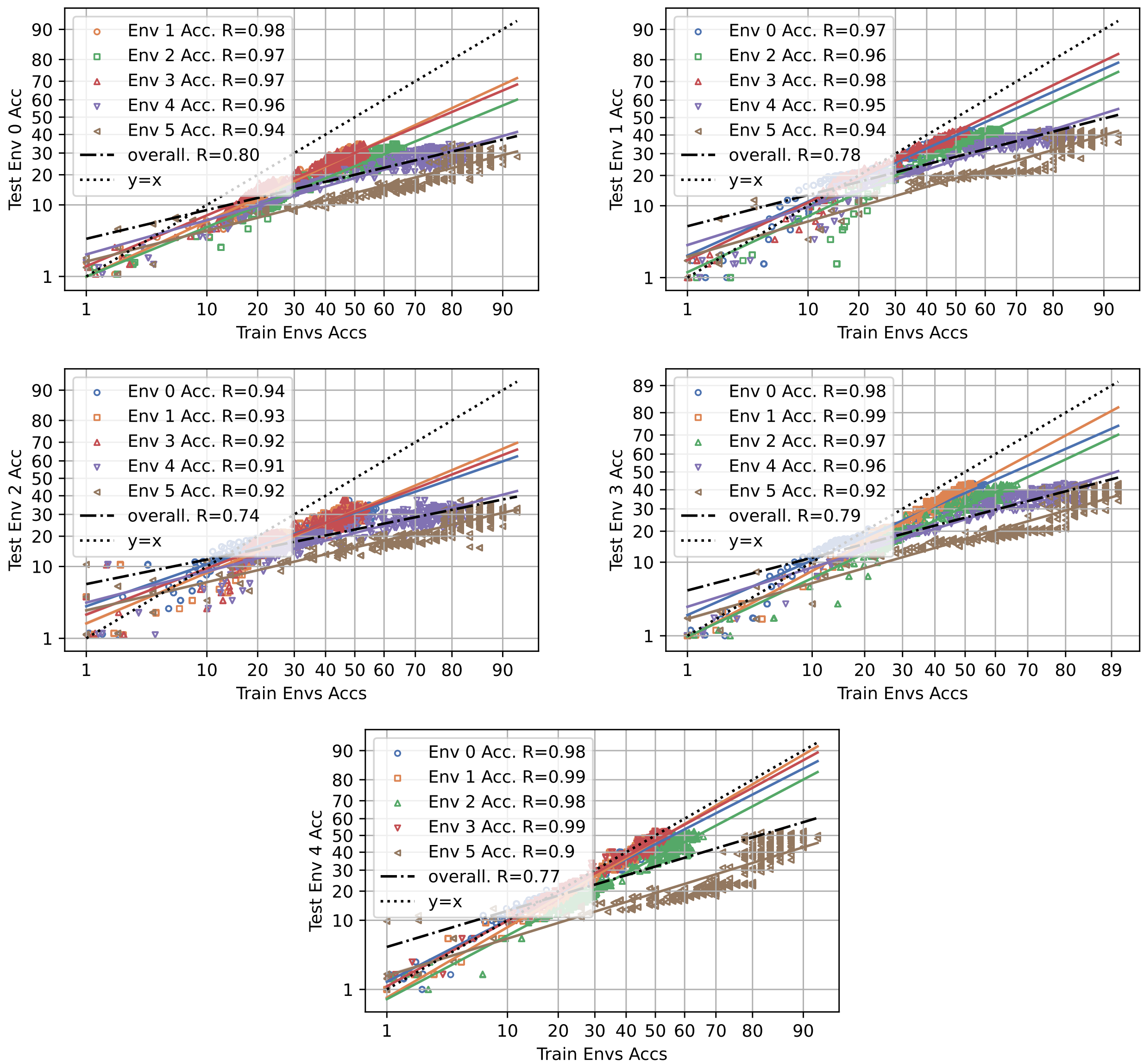


Figure 25: FMoW: Train Env Accuracy vs. Test Env Accuracy.

C.9 Waterbirds

Waterbirds (Sagawa et al., 2019; Koh et al., 2021). The Waterbirds dataset is a modification of the CUB dataset (Welinder et al., 2010) constructed to induce a subpopulation shift in the association between bird type $b \in \{waterbird, landbird\}$ in the foreground and the image background $c \in \{water, land\}$. Waterbird labels are assigned to seabirds and waterfowl; all other bird types are labeled as landbirds. Image backgrounds are obtained from the Places dataset (Zhou et al., 2016) and subset to include water backgrounds (categories: ocean or natural lake) and land backgrounds (categories: bamboo forest or broadleaf forest). The training set consists of 95% of all waterbirds with a water background and the remaining 5% with a land background. Similarly, 95% of all landbirds are displayed on a land background with the remaining 5% on a water background. The validation and test sets include an equal distribution of waterbirds and landbirds on each background. This dataset consists of 11,788 examples of size (3 x 224 x 224).

Experimental Details. Environment 0 consists of the full training split from the Waterbirds dataset and Environment 1 is a concatenation of the validation and test splits from the Waterbirds dataset (Sagawa et al., 2019). In addition to plotting the average ID vs. OOD accuracies for each test environment, we also include average ID vs. group-specific accuracies and pairwise combinations of group-specific ID vs. group-specific OOD accuracy.

Discussion. In our evaluation of average ID and OOD performance, we find that the Waterbirds dataset does not strongly represent worst-case shifts. However, plotting group-specific OOD accuracies, we find no linear correlation for Environment 1 group ($y = 0, a = 1$), representing a degradation in worst-group accuracy (WGA) with improvements to Environment 0 ID average accuracy. We examine the pairwise group-specific accuracies for each environment and generally find that OOD group accuracy positively correlates with ID accuracy for groups sharing the same label. Since Waterbirds groups are defined by label (bird), attribute (background) pairs, this result is consistent with the studies that suggest worst-class accuracy (WCA) is a good proxy for WGA when group membership is unknown (Yang et al., 2023). Similarly, OOD group accuracy negatively correlates with ID accuracy for groups of the opposite label, demonstrating the trade-off between majority and minority group/class performance under subpopulation shifts.

Table 23: WILDSWaterbirds average ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	Env 1 acc	0.83	0.35	0.92	0.00	0.01
Env 1 acc	Env 0 acc	0.47	0.16	0.69	0.00	0.02

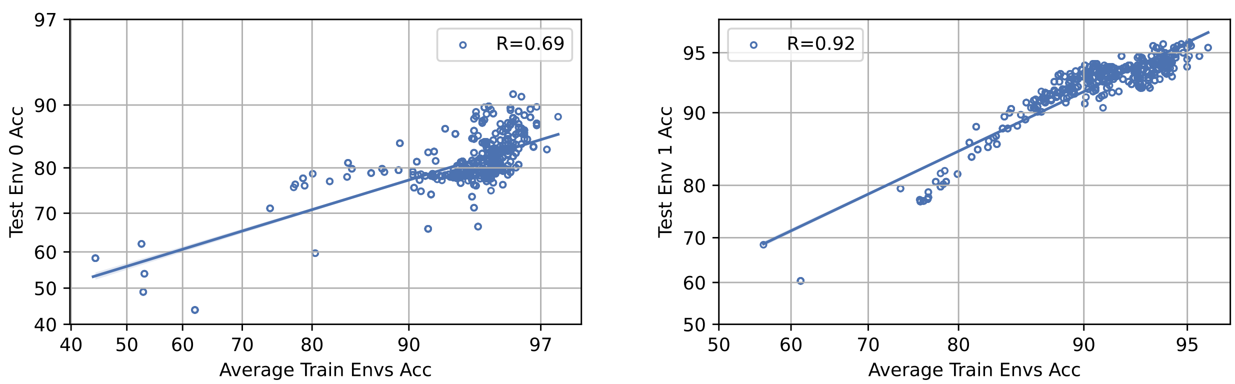


Figure 26: Waterbirds average ID vs. average OOD accuracies

Table 24: Waterbirds average ID vs. group-specific OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 avg acc	Env 1 y=0,a=0 acc	-0.13	1.58	-0.13	0.00	0.03
Env 0 avg acc	Env 1 y=0,a=1 acc	0.00	1.32	0.00	0.95	0.03
Env 0 avg acc	Env 1 y=1,a=0 acc	0.22	1.17	0.84	0.00	0.00
Env 0 avg acc	Env 1 y=1,a=1 acc	0.32	1.13	0.81	0.00	0.01
Env 1 avg acc	Env 0 y=0,a=0 acc	0.70	-0.02	0.65	0.00	0.03
Env 1 avg acc	Env 0 y=0,a=1 acc	-0.07	1.61	-0.11	0.00	0.02
Env 1 avg acc	Env 0 y=1,a=0 acc	0.27	1.63	0.56	0.00	0.01
Env 1 avg acc	Env 0 y=1,a=1 acc	0.30	1.21	0.60	0.00	0.01

Table 25: Waterbirds pairwise group-specific ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 y=0,a=0 acc	Env 1 y=0,a=0 acc	0.76	0.58	0.86	0.00	0.01
Env 0 y=0,a=1 acc	Env 1 y=0,a=0 acc	0.72	0.73	0.73	0.00	0.02
Env 0 y=1,a=0 acc	Env 1 y=0,a=0 acc	-0.15	2.19	-0.43	0.00	0.01
Env 0 y=1,a=1 acc	Env 1 y=0,a=0 acc	-0.13	2.17	-0.36	0.00	0.01
Env 0 y=0,a=0 acc	Env 1 y=0,a=1 acc	0.45	1.05	0.42	0.00	0.03
Env 0 y=0,a=1 acc	Env 1 y=0,a=1 acc	0.65	0.72	0.54	0.00	0.03
Env 0 y=1,a=0 acc	Env 1 y=0,a=1 acc	-0.03	1.96	-0.06	0.08	0.01
Env 0 y=1,a=1 acc	Env 1 y=0,a=1 acc	-0.05	1.99	-0.10	0.00	0.01
Env 0 y=0,a=0 acc	Env 1 y=1,a=0 acc	-1.16	3.05	-0.33	0.00	0.10
Env 0 y=0,a=1 acc	Env 1 y=1,a=0 acc	-0.96	2.58	-0.25	0.00	0.12
Env 0 y=1,a=0 acc	Env 1 y=1,a=0 acc	1.24	0.13	0.93	0.00	0.02
Env 0 y=1,a=1 acc	Env 1 y=1,a=0 acc	1.34	0.12	0.92	0.00	0.02
Env 0 y=0,a=0 acc	Env 1 y=1,a=1 acc	-0.70	2.02	-0.32	0.00	0.07
Env 0 y=0,a=1 acc	Env 1 y=1,a=1 acc	-0.87	2.24	-0.34	0.00	0.08
Env 0 y=1,a=0 acc	Env 1 y=1,a=1 acc	0.84	0.19	0.95	0.00	0.01
Env 0 y=1,a=1 acc	Env 1 y=1,a=1 acc	0.92	0.20	0.97	0.00	0.01
Env 1 y=0,a=0 acc	Env 0 y=0,a=0 acc	0.77	0.55	0.96	0.00	0.01
Env 1 y=0,a=1 acc	Env 0 y=0,a=0 acc	0.16	2.18	0.35	0.00	0.01
Env 1 y=1,a=0 acc	Env 0 y=0,a=0 acc	-0.07	2.23	-0.18	0.00	0.01
Env 1 y=1,a=1 acc	Env 0 y=0,a=0 acc	0.02	2.26	0.05	0.10	0.01
Env 1 y=0,a=0 acc	Env 0 y=0,a=1 acc	0.54	-0.49	0.39	0.00	0.04
Env 1 y=0,a=1 acc	Env 0 y=0,a=1 acc	0.73	0.23	0.88	0.00	0.01
Env 1 y=1,a=0 acc	Env 0 y=0,a=1 acc	-0.13	0.61	-0.23	0.00	0.02
Env 1 y=1,a=1 acc	Env 0 y=0,a=1 acc	-0.50	1.25	-0.69	0.00	0.02
Env 1 y=0,a=0 acc	Env 0 y=1,a=0 acc	-0.33	0.50	-0.18	0.00	0.05
Env 1 y=0,a=1 acc	Env 0 y=1,a=0 acc	-0.40	-0.01	-0.32	0.00	0.04
Env 1 y=1,a=0 acc	Env 0 y=1,a=0 acc	0.61	0.17	0.84	0.00	0.01
Env 1 y=1,a=1 acc	Env 0 y=1,a=0 acc	0.88	-1.21	0.83	0.00	0.02
Env 1 y=0,a=0 acc	Env 0 y=1,a=1 acc	0.05	1.08	0.03	0.38	0.05
Env 1 y=0,a=1 acc	Env 0 y=1,a=1 acc	-0.50	1.50	-0.47	0.00	0.03
Env 1 y=1,a=0 acc	Env 0 y=1,a=1 acc	0.44	1.44	0.60	0.00	0.02
Env 1 y=1,a=1 acc	Env 0 y=1,a=1 acc	0.94	0.13	0.96	0.00	0.01

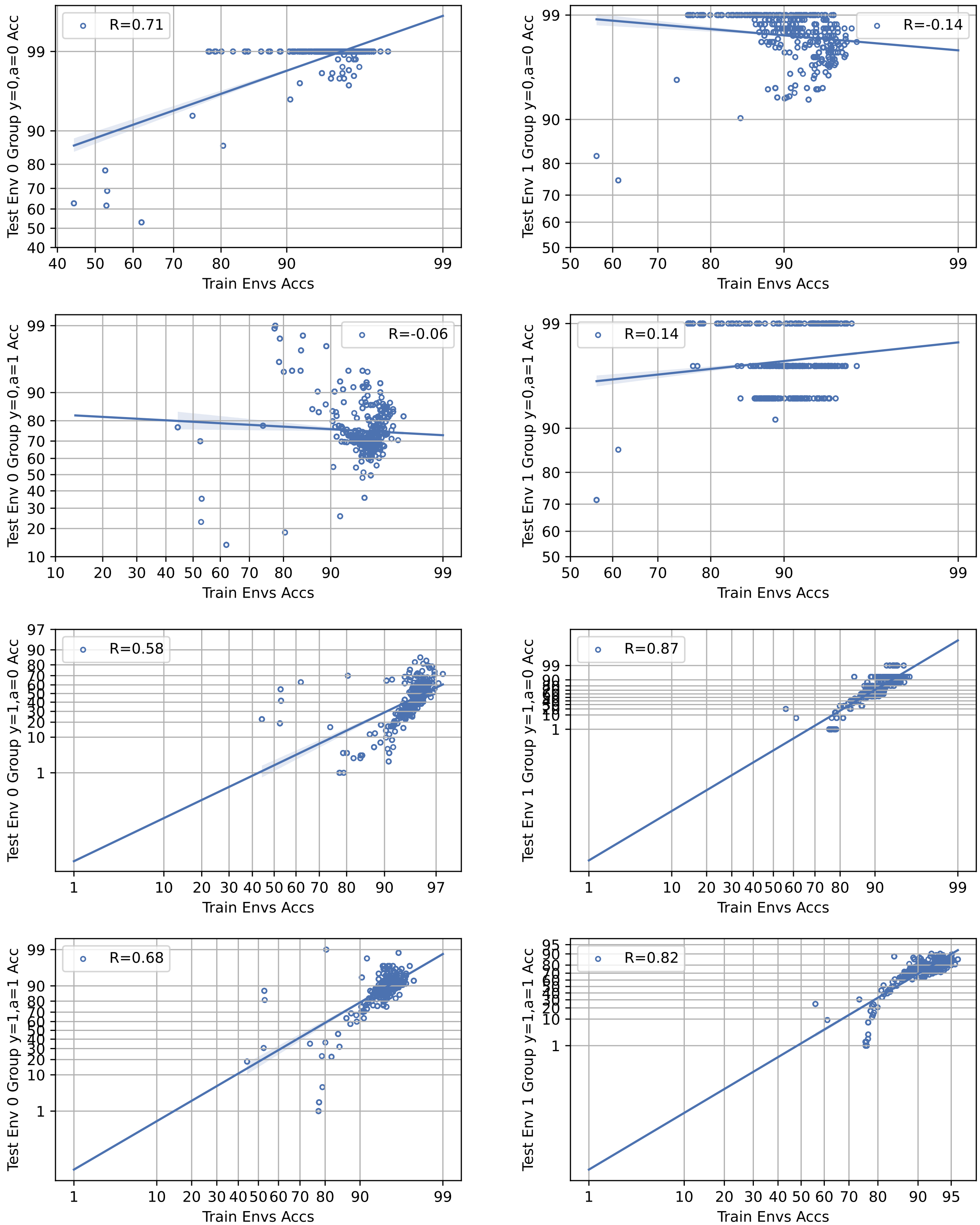


Figure 27: Waterbirds average ID vs. group-specific OOD accuracies

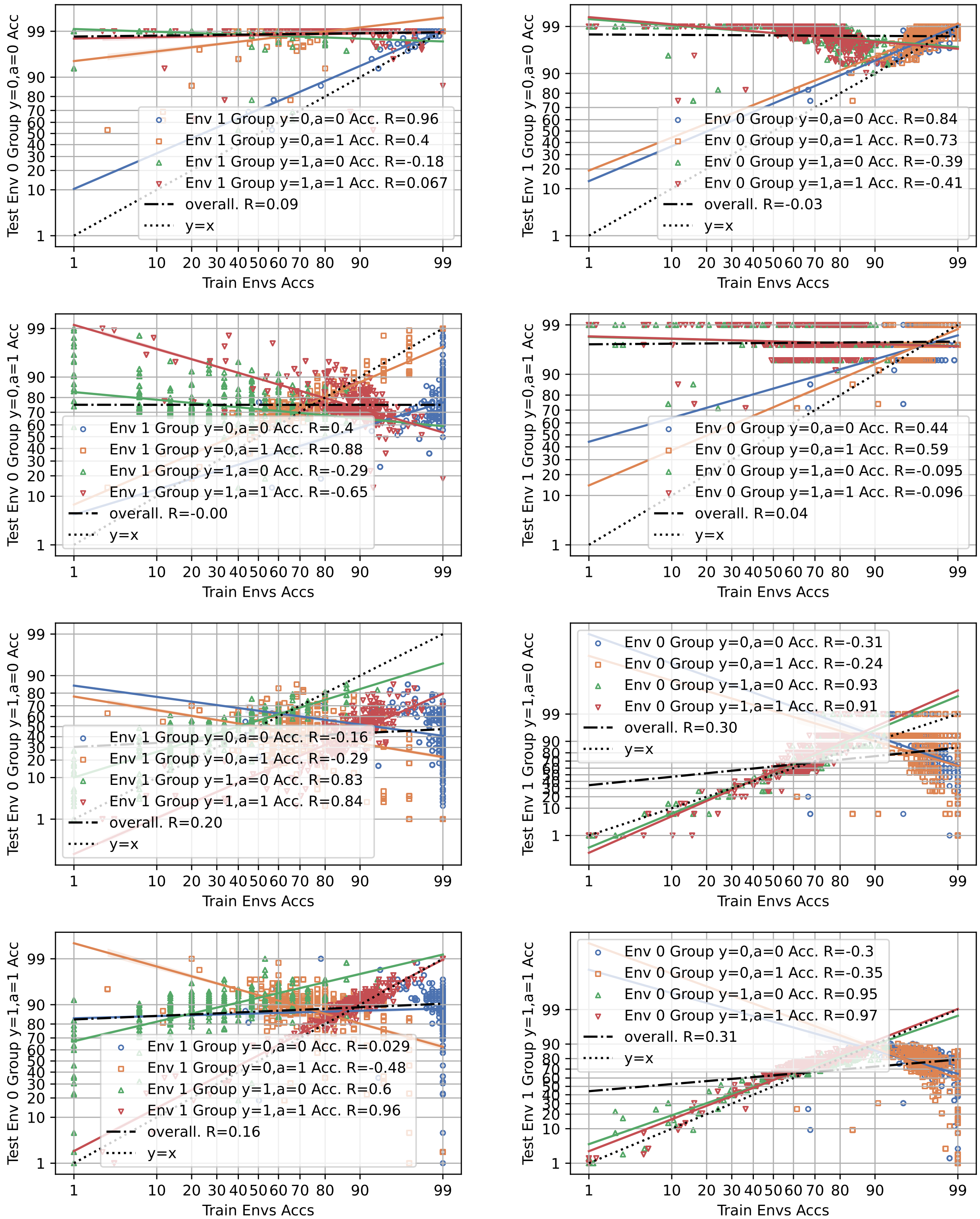


Figure 28: Waterbirds pairwise group-specific accuracies

C.10 CivilComments

CivilComments (Koh et al., 2021) A dataset comprised of multiple subpopulations, which correspond to different demographic identities. The domain d is a multi-dimensional binary vector with 8 dimensions, each corresponding to whether a given comment mentions one of the 8 demographic identities: *male, female, LGBTQ, Christian, Muslim, other religions, Black*, and *White*. For each of the 8 identities, we form 2 groups corresponding to the toxicity label, generating a total of 16 groups (e.g., male_toxic, male_non-toxic, etc.). By construction, since each comment may belong to more than one group, this experimental procedure differs slightly from the standard subpopulation shift framework discussed in this work. Regardless, the experimental results with and without group overlaps display similar accuracy on the line and accuracy on the inverse line patterns.

Experimental Details. We leverage both BERT and DistilBERT architectures for the CivilComments dataset.

Discussion. The CivilComments dataset exhibits both an accuracy on a line and an accuracy on the inverse line phenomenon across all test environments in the Leave-One-Domain-Out procedure. The spurious correlation, which involves the presence or absence of 8 different demographic identities (e.g., male, white, Christian, Muslim, etc.), is rather strongly correlated with the toxicity label. Empirically, the correlation coefficient (R) values range from as low as -0.47 to as high as 0.43. There appear to be many ID/OOD splits that can be derived from this dataset for benchmarking domain generalization.

Table 26: CivilComments ID vs. OOD properties.

OOD	slope	intercept	Pearson R	p-value	standard error
Env 0 acc	0.36	-0.29	0.28	0.00	0.04
Env 1 acc	-0.49	0.16	-0.47	0.00	0.03
Env 2 acc	0.39	-0.17	0.28	0.00	0.04
Env 3 acc	-0.49	0.12	-0.42	0.00	0.03
Env 4 acc	0.58	0.20	0.43	0.00	0.04
Env 5 acc	-0.54	0.00	-0.43	0.00	0.04
Env 6 acc	0.46	-0.04	0.30	0.00	0.05
Env 7 acc	-0.50	0.16	-0.43	0.00	0.03
Env 8 acc	0.49	0.03	0.36	0.00	0.04
Env 9 acc	-0.49	0.06	-0.41	0.00	0.03
Env 10 acc	0.53	0.17	0.34	0.00	0.05
Env 11 acc	-0.57	-0.09	-0.46	0.00	0.03
Env 12 acc	0.46	0.05	0.30	0.00	0.05
Env 13 acc	-0.54	-0.04	-0.45	0.00	0.03
Env 14 acc	0.41	-0.10	0.29	0.00	0.04
Env 15 acc	-0.52	0.06	-0.43	0.00	0.04

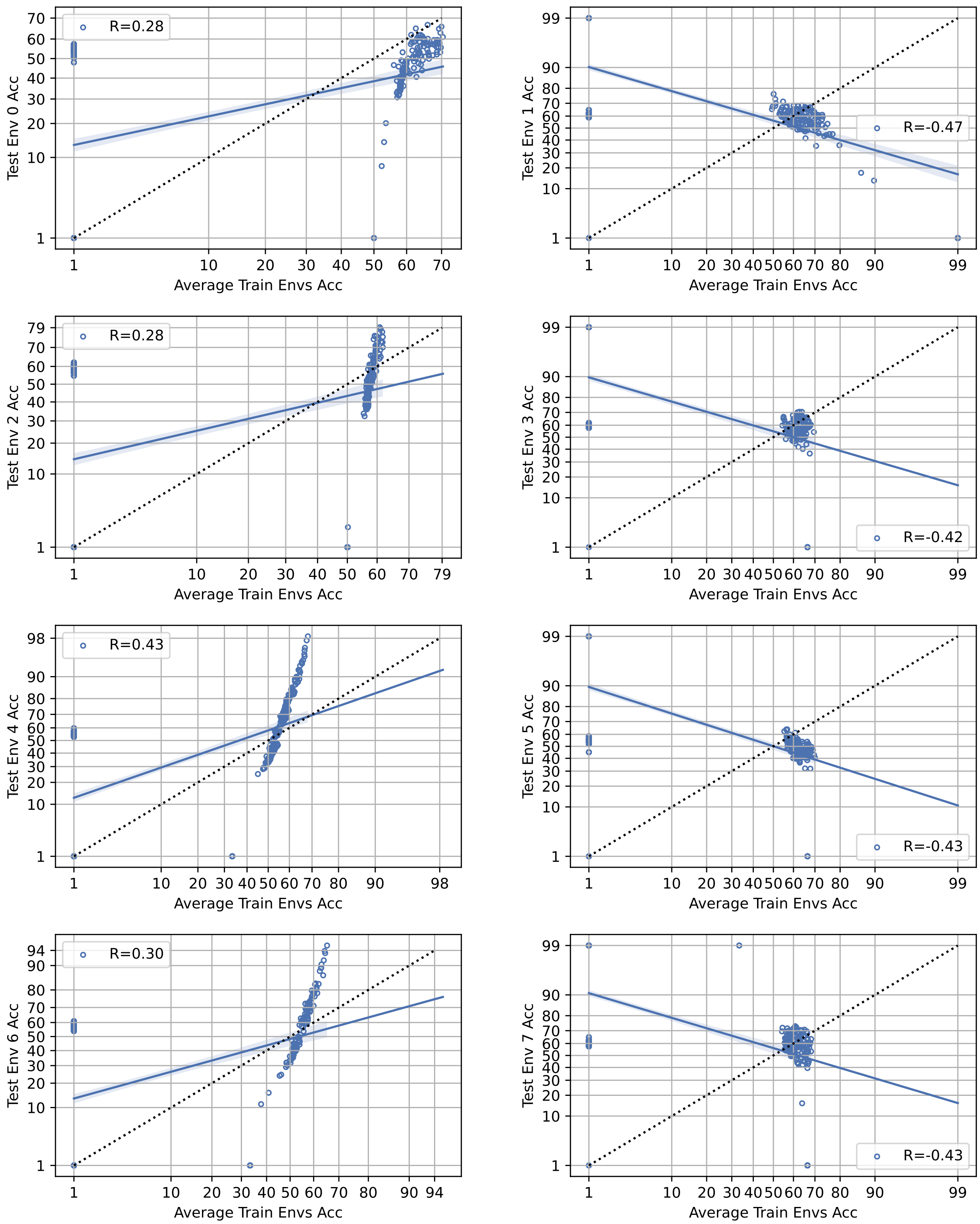


Figure 29: CivilComments: Average train Env Accuracy vs. Test Env Accuracy.

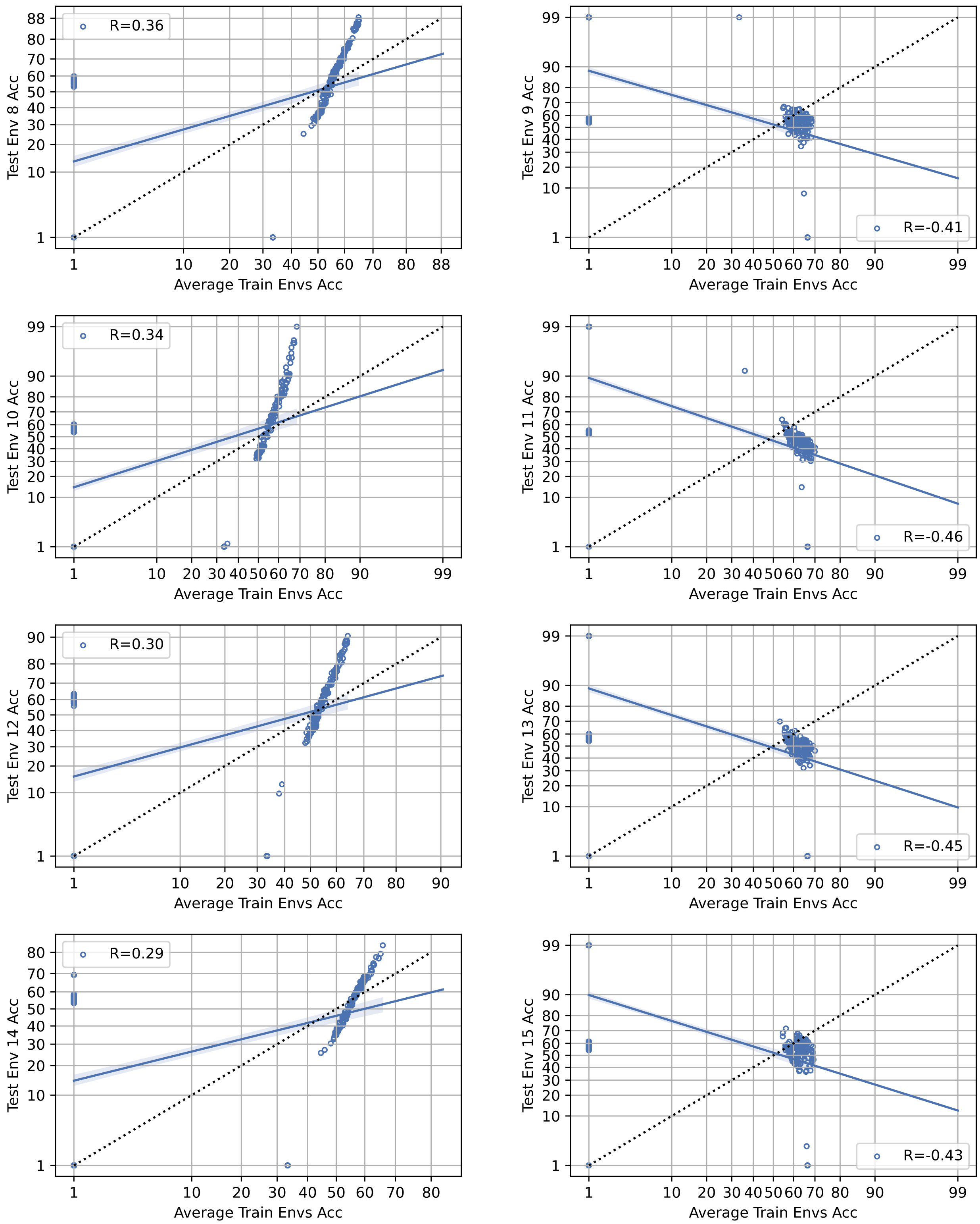


Figure 30: CivilComments: Average train Env Accuracy vs. Test Env Accuracy.

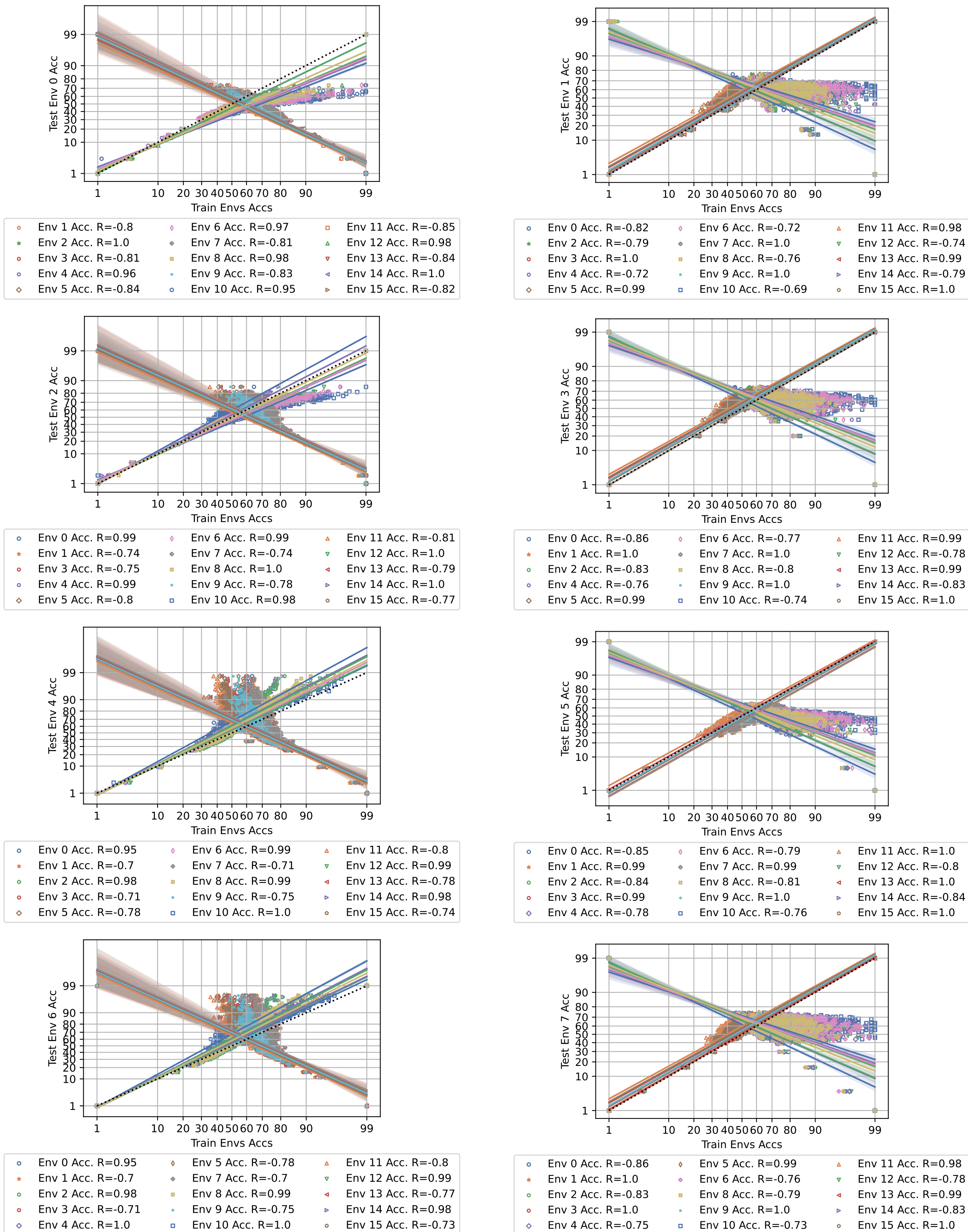


Figure 31: CivilComments: Train Env Accuracy vs. Test Env Accuracy.

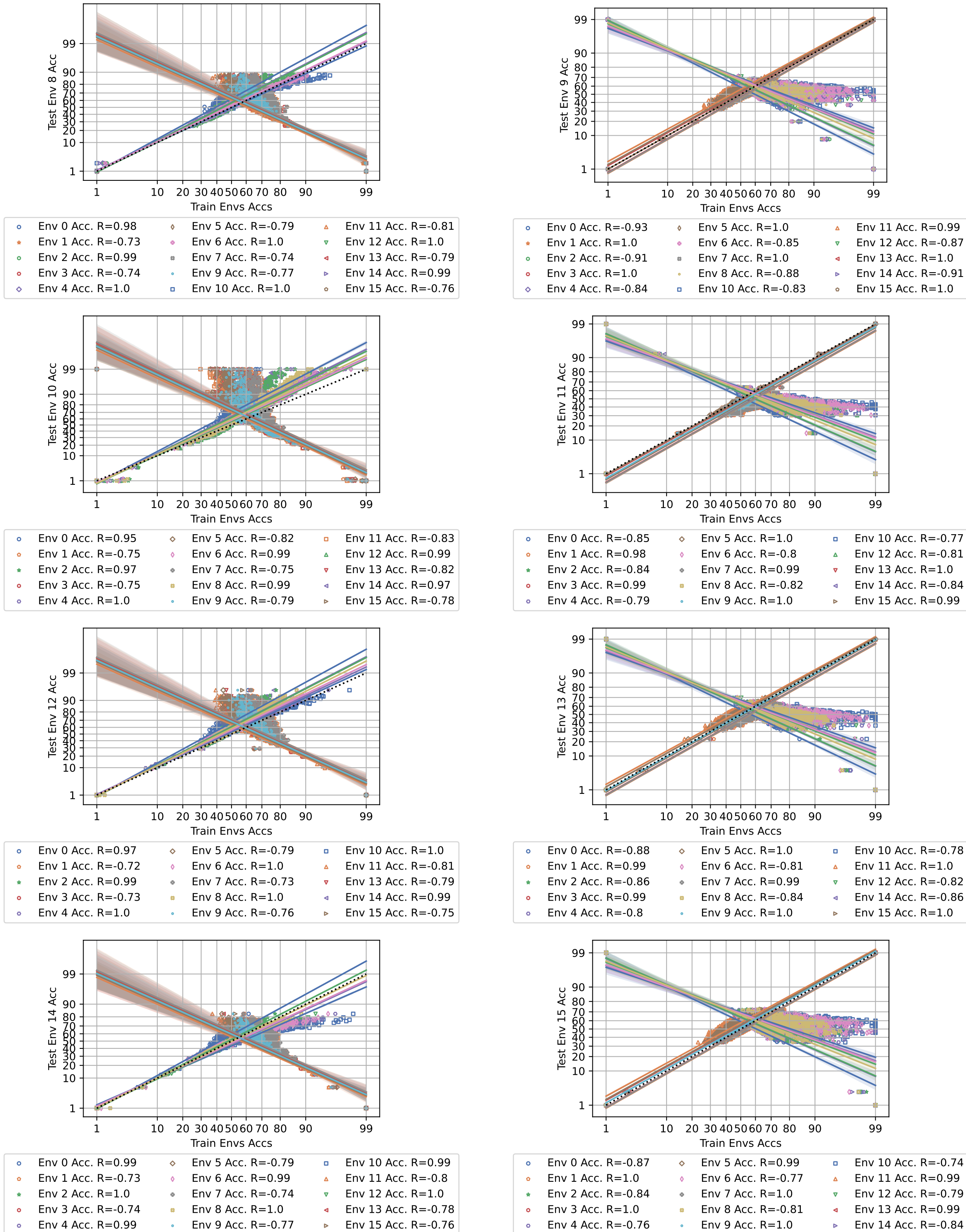


Figure 32: CivilComments: Train Env Accuracy vs. Test Env Accuracy.

D Benchmarking Causal Representation Learning

Models that incorporate or learn structural knowledge of the domains they are applied to have been shown to be more efficient and generalize better across different settings (Parascandolo et al., 2020; Sanchez-Gonzalez et al., 2020; Gondal et al., 2019; Goyal et al., 2019; Battaglia et al., 2016; Bapst et al., 2019; Makar et al., 2022; Zheng & Makar, 2022; Wang et al., 2022a). An example of such a structure is the principle of *independent causal mechanisms* (Haavelmo, 1944; Aldrich, 1989; Hoover, 1990; Pearl, 2009; Schölkopf et al., 2012; Janzing et al., 2012; Peters et al., 2016; Schölkopf et al., 2021), which posits that the generative process of a system's variables consists of autonomous components, or mechanisms, that operate independently and do not inform one another. This implies that the conditional distribution of each variable, given its causes (mechanisms), is independent of the other variables and mechanisms (Peters et al., 2017). Learning causal representations is an active area of research (Schölkopf et al., 2021). Datasets for causal representation learning are primarily (semi)parametric where (some) causal variables are known and potentially intervenable (Von Kügelgen et al., 2021; Lippe et al., 2022a;b; Ahmed et al., 2020; Liu et al., 2023b). Then, success is assessed by how well learned disentangled representations (mechanisms) explain outcome variance, using R^2 or MCC (Mathew's Correlation Coefficient) (Lopez et al., 2023). However, the task of causal representation learning with complex datasets with limited knowledge or control over generative mechanisms remains a challenge, especially without requiring most (or at least some) relevant causal variables to be directly observed Lopez-Paz et al. (2017)—we identify that benchmarking causal representation learning in this setting is also challenging.

Causal representation learning is closely tied to domain generalization, which aims to learn representations from multiple observed domains that give predictors whose performance is invariant to new domains (new data distributions). Many works in domain generalization (Arjovsky et al., 2019; Salaudeen & Koyejo, 2022; 2024; Mahajan et al., 2021; Liu et al., 2021; Lv et al., 2022; Chen et al., 2022; Eastwood et al., 2022). have been motivated by the principle of independent causal predictors, which aims to identify causal predictors from observational data by searching for feature sets that maintain stable (invariant) predictive accuracy across interventional distributions Peters et al. (2016); Heinze-Deml et al. (2018); Arjovsky et al. (2019). Additionally, more recent work motivates learning causal representations from multiple datasets arising from unknown interventions (von Kügelgen et al., 2024).

Thus, one may naively consider domain generalization as a proxy task to benchmark causal representation learning in more complex settings. This work studies when performance on a domain generalization task is informative of the causal representation learning task. Specifically, when benchmarking a set of models, including a disentangled causal model, based on transfer accuracy, our results on evaluating domain generalization apply, where non-causal correlations are spurious (Salaudeen et al., 2024).