DeconDTN-Toolkit: A Library for Evaluation and Enhancement of Robustness to Provenance Shift

Yongsen Tan[†*] TANYS@UW.EDU
Zhecheng Sheng[‡*] SHENG136@UMN.EDU
Xiruo Ding[†*] XIRUOD@UW.EDU
Serguei V. S. Pakhomov[‡] PAKH0002@UMN.EDU
Trevor Cohen[†] COHENTA@UW.EDU
[†] *University of Washington,* [‡] *University of Minnesota*

Abstract

Despite the burgeoning body of work on distribution shifts, provenance shift—where the relationship between data source and label changes at deployment—remains poorly understood and under-addressed. In this paper, we establish a formal connection between provenance shift, counterfactual invariance, and invariant learning to derive a learning objective for robustness. We then introduce DeconDTN-Toolkit, a specialized evaluation and remediation suite designed to simulate provenance shifts of varying degrees while maintaining the training protocol and the infrastructure of existing benchmarks. We reveal the vulnerability of Empirical Risk Minimization under provenance shift, introduce a robust out-of-distribution performance indicator, and conduct a comprehensive evaluation on existing algorithms. Our work provides both the theoretical grounding and the practical tools necessary to characterize the problem of confounding by provenance, and implementations of methods to mitigate it.

Data and Code Availability This paper uses five publicly available datasets: **SHAC** (Lybarger et al., 2021) is available upon request, **MIMIC-III** (Johnson et al., 2016) is available through PhysioNet[1], **HateSpeech** (Vidgen et al., 2021; De Gibert et al., 2018) is available on the investigators' GitHub repository[2][3], **MultiNLI** (Williams et al., 2018) via Huggingface[4], and **Civilcomments** (Borkan et al., 2019) is available through Kaggle[5]. We include a detailed data access and usage instructions in DeconDTN-Toolkit (https://github.com/LinguisticAnomalies/DeconDTN-toolkit).

Author Contributions Y.T. developed the theoretical framework, led the toolkit development, performed the experiments, analyzed the data, and wrote the manuscript. Z.S. and X.D. developed the initial toolkit, aided in toolkit refactorization, conceived the evaluation framework, and contributed to the interpretation of the results. T.C. and S.P. conceived the study and were in charge of overall direction and planning. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Institutional Review Board (IRB) This study was approved by the University of Washington IRB Committee (protocol ID STUDY00015108).

1. Introduction

The success of modern machine learning systems relies critically on the quality and quantity of their training data. A common practice to scale the data quantity and reduce generalization error in natural language processing (NLP) is to mix multiple data sources for a unified task (Laparra et al., 2020). For example, the pre-training corpus for masked language modeling and next sequence pre-

[*] These authors contributed equally
1. https://physionet.org/content/mimiciii/1.4
2. https://github.com/Vicomtech/hate-speech-dataset
3. https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset
4. https://huggingface.co/datasets/nyu-mll/multi_nli
5. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

diction tasks of BERT were BooksCorpus and English Wikipedia (Devlin et al., 2019), and more recent efforts have sought to develop and validate clinical NLP algorithms collaboratively within an established data-sharing network (Wang et al., 2025). Counter-intuitively, increasing the number of data sources can harm systems' performance by introducing spurious correlations to the training distribution in real-world settings (Guo et al., 2021; Compton et al., 2023; Shen et al., 2024). A mechanism through which machine learning models might internalize such spurious correlations was revealed by the "Name That Dataset" experiment, in which a neural network was trained to identify the source of an image sampled from a variety of datasets (Torralba and Efros, 2011). Notably, the dataset an image was drawn from could be identified with high accuracy in 2011, using deep learning methods (Krizhevsky et al., 2012; Liu and He, 2024b). Modern neural networks, therefore, can harness source-specific features for label prediction (Geirhos et al., 2020), resulting in flawed systems that encode spurious correlations between data sources and labels. For example, in the context of a multi-site set of clinical notes, a system may learn to associate a characteristic acronym with the prevalence of an outcome of interest at one location, which would lead to inaccurate predictions at the point of deployment if the differences in prevalence across sites do not match those in the training set (Howell et al., 2020; Koh et al., 2021; Yang et al., 2023). This algorithmic bias has been referred to as *confounding by provenance* in previous work (Ding et al., 2023).

There is a burgeoning body of work on addressing robustness to test-time distribution shifts, including domain generalization (Zhou et al., 2023), label shift (Lipton et al., 2018), and subpopulation shift (Koh et al., 2021; Yang et al., 2023). However, this prior work does not focus on provenance shift, and the utility of methods developed to address these distribution shifts in the context of confounding by provenance has yet to be established. In addition, although recent theoretical and algorithmic frameworks have been proposed to account for observed confounders in machine learning (Landeiro and Culotta, 2016, 2018; Veitch et al., 2021; Schrouff et al., 2024; Zhang et al., 2024), there is an unmet need for a unified toolkit to facilitate the systematic evaluation of robustness to provenance shift, and evaluate approaches with the potential to address it.

In this work, we provide a systematic approach to the study of robustness to provenance shift in the context of multiple data sources. We first introduce the problem of provenance shift and establish a formal connection between it and the related domains of counterfactual invariance and invariant learning by deriving a robustness learning objective under provenance shift (Section 2). To facilitate empirical study of provenance shift, we introduce the DeconDTN-Toolkit (Deconfounding Deep Transformer Networks Toolkit), which can synthetically introduce varying degrees of provenance shift between train and test time and provides a standardized interface to a range of established and recently-developed approaches (Section 3). Using DeconDTN-Toolkit, we investigate the learning dynamics of Empirical Risk Minimization (ERM), identify a predictor of out-of-distribution performance, and test for robustness a set of invariant learning algorithms under increasing degrees of provenance shift (Sections 4 and 5). Our contributions are summarized as follows:

1. DeconDTN-Toolkit, a specialized and systematic suite to simulate provenance shifts of varying degrees, evaluate the robustness of text classification models, and remediate vulnerability to provenance shift using invariant learning algorithms.

2. We formalize provenance shift, derive a risk-invariant objective under specific causal graphs, and, through a comprehensive empirical study with DeconDTN-Toolkit, demonstrate that standard ERM and oracle selection fail under such shifts.

3. A comprehensive benchmarking on invariant learning algorithms under provenance shift with aligned evaluation protocol to previous study. We show that removing the correlation between provenance Z and outcome Y in the training distribution remains strong baseline performance under provenance shift.

2. Problem Setting

To formalize the problem of provenance shift, we consider a supervised learning task where the goal is to learn a predictor $f : \mathcal{X} \to \mathcal{Y}$ using a labeled dataset $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^{n}$, where Z denotes observed confounders. In the problem of provenance shift, Z denotes the provenance of a subset of data within $\mathcal{D}$. Given a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ defining the

learning objective, supervised learning seeks a predictor f that minimizes the risk $\mathbb{E}_{(x,y)\sim P}[\mathcal{L}(f(x), y)]$. We assume that samples are independent and identically distributed (i.i.d.) according to the causal graph $\mathcal{G} : X \leftarrow Z \rightarrow Y$.

This work investigates the problem of learning a predictor f that is counterfactually invariant to the confounder Z (e.g. $f(X(z)) = f(X(z'))$), thereby achieving robustness to provenance shift at test time. In the following sections, we first define, decompose, and characterize provenance shift (Section 2.1). Then, we connect counterfactual invariance to provenance shift, derive a risk-invariant learning objective under specific causal graphs, and propose practical approaches to achieve the objective (Section 2.2).

2.1. Provenance Shift

Provenance shift refers to a test-time change in the correlation between the provenance Z and the outcome Y[6]. We assume that the underlying causal graph $\mathcal{G}$ remains invariant across the training distribution P^{tr} and the test distribution P^{te}.

Definition 1 (Provenance Shift) *Let Z be an observed confounder (i.e., common cause) of the outcome Y and the observed variables X in the causal graph $\mathcal{G} : X \leftarrow Z \rightarrow Y$. Provenance shift occurs when $P^{tr}(Y \mid Z) \neq P^{te}(Y \mid Z)$, while $\mathcal{G}$ remains unchanged.*

Example 1 (Substance) *Consider detecting substance abuse mentions Y in clinical notes X from two sites $Z \in \{z_1, z_2\}$. In the training environment, site z_1 utilizes purposive sampling, resulting in a high prevalence of mentions $P^{tr}(Y \mid Z = z_1)$. During deployment, site z_1 switches to standard clinical flow, where the prevalence $P^{te}(Y \mid Z = z_1)$ is significantly lower and no longer artificially elevated. Even if site z_2 remains stable, the system may overestimate substance abuse for notes from z_1.*

Example 2 (Goals-of-Care) *Consider detecting goals-of-care discussions, a prerequisite to goal-concordant end-of-life care, from clinical notes. Let $Z \in \{z_1, z_2\}$ denote the note provenance, where z_1 corresponds to palliative care specialist notes and z_2 to notes from a random sample of emergency admissions. In the training environment, a convenience sample may overrepresent palliative care notes, yielding a higher prevalence of goals-of-care discussions $P^{tr}(Y \mid Z = z_1)$. During deployment, however, site z_1 may shift to random universal screening, causing $P^{te}(Y \mid Z = z_1)$ to drop significantly as the "purposive" focus on high-risk patients is removed. Conversely, site z_2 might open a "New Wing" dedicated to intensive care planning, causing the prevalence $P^{te}(Y \mid Z = z_2)$ to rise relative to the original general admission baseline $P^{tr}(Y \mid Z = z_2)$. A system deployted across Z can overestimate discussions for site z_1 and underestimate them for site z_2.*

From a dataset bias perspective — wherein neural networks can learn generalizable and transferable features of provenance (Liu and He, 2024a), we decompose the discriminative model that follows the causal graph $\mathcal{G}$ into two components: the inference mechanism (the capacity to map spurious features back to their latent causes) and the provenance mechanism (the observed outcome prevalence conditioned on provenance):

Lemma 2 *(Prediction Decomposition) Let $X_Y^{\perp}$ be Y-invariant components of the input features, such that $X_Y^{\perp}(y) = X_Y^{\perp}(y')$. A discriminative model under the causal graph $\mathcal{G} : X \leftarrow Z \rightarrow Y$ can be decomposed into two components: the inference and the provenance mechanism.*

$$P(Y \mid X_Y^{\perp}) = \int \underbrace{P(Y \mid Z)}_{provenance} \underbrace{P(Z \mid X_Y^{\perp})}_{inference} dZ \quad (1)$$

Detailed proofs are provided in A.1. The instability of f under provenance shift stems from features in X that serve as proxies for Z (i.e., inference mechanism). Intuitively, Z serves as an intermediary between $X_Y^{\perp}$ and Y, and the prediction shifts if the provenance-specific label distribution changes (Definition 1) while the inference mechanism is generalizable for predictor f at test-time. Because Z is predictive of Y in the training distribution, the model may rely on these proxies $X_Y^{\perp}$ despite the absence of a direct causal link between the proxy features $X_Y^{\perp}$ and the outcome Y.

We compare provenance shift with a set of related concepts in Table 1. Notably, provenance shift does not assume a distribution shift in $P(Y)$ or $P(Z)$, nor does it require the model f to generalize to unseen values of Z. Furthermore, performance degradation under provenance shift does not strictly require attribute imbalance, class imbalance, or attribute generalization; rather, it manifests as a specific form of

6. Provenance shift is a form of *confounding shift* as defined by Landeiro and Culotta (2018).

Table 1: Distribution Shifts in Supervised Learning. L^z and U^z denote the labeled and unlabeled distribution from domain z, respectively

Concept	Train Inputs	Test Inputs	Test-Time Invariance	Test-Time Change
Label Shift	L^1	U^1	$P^{tr}(Y \mid X) = P^{te}(Y \mid X)$	$P^{tr}(Y) \neq P^{te}(Y)$
Domain Generalization	$L^1, \ldots, L^{z^{tr}}$	$U^{z^{tr+1}}$	$P^{tr}(Y \mid X) = P^{te}(Y \mid X)$	$P^{tr}(X) \neq P^{te}(X)$
Spurious Correlation	$L^1, \ldots, L^{z^{tr}}$	$U^1, \ldots, U^{z^{tr}}$	$P^{tr}(Y \mid X_Z^\perp) = P^{te}(Y \mid X_Z^\perp)$	$P^{tr}(Y \mid Z) \neq P^{te}(Y \mid Z)$
Subpopulation Shift	$L^1, \ldots, L^{z^{tr}}$	$U^1, \ldots, U^{z^{tr+n}}$	$P^{tr}(X, Y \mid Z) = P^{te}(X, Y \mid Z)$	$P^{tr}(Z) \neq P^{te}(Z)$
Provenance Shift	$L^1, \ldots, L^{z^{tr}}$	$U^1, \ldots, U^{z^{tr}}$	$\mathcal{G}^{tr} = \mathcal{G}^{te}, X \leftarrow Z \rightarrow Y$	$P^{tr}(Y \mid Z) \neq P^{te}(Y \mid Z)$

spurious correlation resulting from the confounding structure $\mathcal{G}$ (Ding et al., 2024).

2.2. Robustness to Provenance Shift

The learning objective of counterfactual invariance is to learn a robust predictor which is counterfactual invariant to the confounder Z:

Definition 3 (Counterfactual Invariance) *(Restated Definition 1 from Veitch et al. (2021)) Denote $X(z)$ as the counterfactual X would have observed had Z been set to z via intervention, leaving all other conditions fixed. A predictor f is counterfactual invariant to Z if $f(X(z)) = f(X(z'))$ almost everywhere, for all $z, z' \in \mathcal{Z}$.*

To derive a learning objective that ensures robustness under provenance shift, we establish a formal connection between counterfactual invariance and provenance shift:

Proposition 4 (Provenance Robustness) *If a predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$ satisfies counterfactual invariance such that $f(X(z)) = f(X(z'))$ for all $z, z' \in \mathcal{Z}$, then the predictor is robust to provenance shift. Specifically, the risk $\mathbb{E}[\mathcal{L}(f(X), Y)]$ remains constant under any intervention on the provenance-specific class distribution $P(Y \mid Z)$, provided $P(Y)$ remains invariant. This robustness holds under both:*

1. *Anti-causal settings $Y \rightarrow X$;*
2. *Causal settings $X \rightarrow Y$: Provided that $Y \perp X \mid \{X_Z^\perp, Z\}$ and the label satisfies counterfactual consistency, $Y(z) = Y(z')$ for all $z, z' \in \mathcal{Z}$.*

Detailed proofs are provided in Appendix A.2. While counterfactuals are often unobservable, we can achieve the necessary conditions for provenance robustness by enforcing the predictor f to satisfy the necessary condition of counterfactual invariance under the causal graph $\mathcal{G}$ in Proposition 4 (Veitch et al., 2021):

$$f(X) \perp Z \mid Y \tag{2}$$

In practice, this learning objective can be achieved by invariant learning, which aims to achieve an invariant prediction $f(X) \perp Z \mid Y$ using a labeled dataset $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^n$ partitioned by attributes $z \in \mathcal{Z}$ (e.g., domains in domain generalization or subgroups in subpopulation shift). To this end, we derive the learning objective of provenance robustness and the practical approach to achieve it. If Z is entirely latent and has no proxy, identifying a shift in $P(Y \mid Z)$ is mathematically intractable. For identifiability, we assume that the confounder Z is observed during training and validation and remains unobserved at test time in the current work.

3. DeconDTN-Toolkit: A Toolkit to Quantify and Improve Robustness to Provenance Shift

DeconDTN-Toolkit is an evaluation suite designed to evaluate model robustness to provenance shift. The *evaluation* component of the toolkit generates test splits with increasing degrees of provenance shift systematically, formalizing the evaluation approach for provenance shift developed by Landeiro and Culotta (2016) following prior work by Ding et al. (2023). The *mitigation* component is built as an extension of DomainBed, the standard domain generalization toolkit for computer vision (Gulrajani and Lopez-Paz, 2020). DeconDTN-Toolkit also inherits the standardized training protocol in DomainBed

and SubpopBench for finding alignment (Gulrajani and Lopez-Paz, 2020; Yang et al., 2023).

The initial release of DeconDTN-Toolkit includes ingestion pipelines for five datasets (Table 2) and implementations of nineteen algorithms, within an extensible infrastructure adapted from DomainBed to support custom datasets and algorithms.

The user interface of DeconDTN-Toolkit is inspired by the TRL library[7], utilizing a centralized `Trainer` class to manage algorithms, datasets, and training configurations. Following DomainBed, we adopt a provenance-balanced loading strategy during training; for any minibatch $\mathcal{B}$, the distribution $P_{\sim\mathcal{B}}(Z)$ is enforced to be uniform.

3.1. Simulation/Evaluation

To facilitate evaluation of robustness to confounding by provenance, DeconDTN-Toolkit can introduce spurious correlations of arbitrary strength between the label Y and provenance Z by specifying the $\{P^{tr}(Y,Z), P^{te}(Y,Z)\}$. If the split sizes $\{|\mathcal{D}^{tr}|, |\mathcal{D}^{te}|\}$ are not explicitly defined, the toolkit employs a greedy subsampling strategy to maximize the feasible sample size. Finally, we ensure that the sampled corpus remains i.i.d., with samples from the same subject strictly partitioned into the same subset to prevent data leakage.

To quantify the degree of shift, we employ a parameter $\alpha \in \mathbb{R}^{|Y|\times|Z|}$ developed by Ding et al. (2023) to measure the correlation between Y and Z:

$$\log \alpha_{ij} = \log \frac{P(Y = i \mid Z = j)}{P(Y = i \mid Z \neq j)} \tag{3}$$

α_{ij} describes the ratio of the conditional distribution of outcome i between provenance j versus others. In binary Y and Z settings, which are the focus of the toolkit currently, we define $\log \alpha := \log \frac{P(Y=1|Z=1)}{P(Y=1|Z=0)} \in \mathbb{R}$. Intuitively, $\log \alpha = 0$ represents a jointly balanced distribution, $\log \alpha > 0$ represents cases in which the prevalence of the primary outcome in provenance $Z = 1$ is greater than another. The sampling-based simulation framework in DeconDTN-Toolkit facilitates the generation of training and testing splits with custom α values. The difference between $\log \alpha$ at training and test time then provides a measure of the magnitude of the provenance shift that has been introduced.

7. https://huggingface.co/docs/trl/en/index

3.2. Algorithms/Mitigation

The initial release of DeconDTN-Toolkit includes nineteen algorithms as follows: (1) Baseline: Empirical Risk Minimization (**ERM**); (2) Sampling: **UpSampling**, **DownSampling** (Japkowicz, 2000); (3) Marginal distribution adjustment: Backdoor Adjustment (**BackDoor**, Landeiro and Culotta (2016)), Marginal Transfer Learning (**MTL**, Blanchard et al. (2021)); (4) Data augmentation: Mixup (**Mixup**, Zhang et al. (2018)), Learning Invariant Predictors with Selective Augmentation (**LISA**, (Yao et al., 2022)), (5) Distribution matching: Deep Correlation Alignment (**CORAL**, Sun and Saenko (2016)), Maximum Mean Discrepancy (**MMD**, Li et al. (2018b)), Optimal Representations for Covariate Shift (**CAD**, Ruan et al. (2021)); (6) Gradient matching: Gradient Matching for Domain Generalization (**Fish**, Shi et al. (2021)); (7) Adversarial training: Domain Adversarial Neural Network (**DANN**, Ganin et al. (2016)), Conditional Domain Adversarial Neural Network (**CDANN**, Li et al. (2018c)); (8) Invariant feature learning: Invariant Risk Minimization (**IRM**, Arjovsky et al. (2020)); (9) Group robust learning: Group Distributionally Robust Optimization (**GroupDRO**, Sagawa* et al. (2019)); (10) Two-stage training: Just Train Twice (**JTT**, Liu et al. (2021)), Deep Feature Reweighting (**DFR**, Kirichenko et al. (2023)), Learning from Failure (**LfF**, Nam et al. (2020)), Dual Filter (**DualFilter**, Sheng et al. (2025)). We included a detailed related work in Appendix B. Algorithm descriptions and their hyperparameters can be found in Appendix C and Table 4, respectively.

3.3. Datasets

The corpora supported by ingestion pipelines in DeconDTN-Toolkit fall into two categories (Table 2):

1. **Source Datasets**: **SHAC** (Lybarger et al., 2021), **MIMIC-Location** (Johnson et al., 2016), and **HateSpeech** (Vidgen et al., 2021; De Gibert et al., 2018). These datasets consist of samples drawn from distinct sources, where *provenance* acts as the confounder.

2. **Attribute Datasets**: **MultiNLI** (Williams et al., 2018), **MIMIC-SubpopBench** (Yang et al., 2023), and **Civilcomments** (Borkan et al., 2019). These datasets comprise sam-

Table 2: Corpora for use in DeconDTN-Toolkit. **Bold** indicate **Source Datasets** and Light indicate Attribute Datasets.

Dataset	Prediction (Y)	Attribute (Z)	Sample size	Adapted from
SHAC	Drug Abuse	Data Source	4,405	Lybarger et al. (2021)
MIMIC-Location	Mortality	Data Source	16,282	Johnson et al. (2016)
HateSpeech	Hate Speech	Data Source	51,847	Vidgen et al. (2021), De Gibert et al. (2018)
MIMIC-SubpopBench	Mortality	Sex	25,880	Yang et al. (2023)
Civilcomments	Hate Speech	Race	447,998	Borkan et al. (2019)
MultiNLI	Entailment	Genre	392,702	Williams et al. (2018)

ples differentiated by attributes, where *attributes* serve as the confounder.

Among the **Source Datasets**, **SHAC** and **MIMIC-Location** permit exploration of provenance shift in the context of clinical NLP, the motivating use case for development of DeconDTN-Toolkit. On account of a dearth of publicly available multi-institutional clinical NLP datasets, we have also included publicly available Attribute Datasets which permit a broader evaluation of robustness to provenance shift and facilitate a reproducible demonstration of the capabilities of DeconDTN-Toolkit. In experiments, we explored the extent to which findings related to provenance shifts in clinical NLP generalize to shifts related to other attributes in clinical and general domain text. To demonstrate the current capabilities of the toolkit, we selected a confounder Z and the outcome Y when multiple options were available. Appendix D provides details on datasets and the variable selection.

4. Experimental Setup

The sections that follow present an evaluation of some of the algorithms implemented within DeconDTN-Toolkit, to demonstrate its capabilities and characterize certain aspects of provenance shift. We detail the experiment in Appendix E.

Distribution Manipulation Following Landeiro and Culotta (2016) and Veitch et al. (2021), we introduced spurious correlations between the label Y and the provenance Z with a training correlation strength of $\log \alpha^{tr} = -0.6$ (Equation (3)) across all experimental settings[8]. To perturb the test distributions under provenance shift, we generate a suite of test splits where $\log \alpha^{te}$ varies linearly from -1 to 1. To isolate the performance impact of subgroup and label shifts, we enforce the constraints $P^{tr}(Y) = P^{te}(Y)$ and $P^{tr}(Z) = P^{te}(Z)$. We further ensure that $P(Y)$ and $P(Z)$ follow uniform distributions to ablate the effects of class and provenance imbalance. We denote the distribution with $\log \alpha^{te} = -0.6$ as the in-distribution test set and $\log \alpha^{te} = 0.6$ as the out-of-distribution (OOD) test set, using performance in the OOD as one measure of robustness to provenance shift.

Hyperparameter Selection Following the training protocol in Gulrajani and Lopez-Paz (2020), we conducted 16 random searches over the joint distribution of hyperparameters for each algorithm and dataset to ensure a fair "best-versus-best" comparison. We list all hyperparameters, their default values, and the joint distribution for random hyperparameter searching in Table 4. We select the optimal hyperparameters based on the validation worst-group accuracy (WGA), consistent with the methodology in Yang et al. (2023).

Model Selection To estimate the stability and standard deviation of each algorithm, we fix the optimal hyperparameters and rerun experiments across 5 distinct random seeds. For each run, we select the model checkpoint that maximizes the worst-group validation accuracy (Yang et al., 2023).

8. $P(Y = 1 \mid Z = 1) : P(Y = 1 \mid Z = 0) \approx 1 : 4$

5. Results

5.1. Learning Dynamics in ERM

We investigate the learning dynamics of ERM using optimal hyperparameters to determine:

1. Whether neural networks rely on shortcuts to artifacts related to provenance (or other confounding variables) during training?

2. Whether a robust ERM solution can be identified via oracle selection on the OOD distribution?

The training progress is standardized to a scale of $[0, 1]$ for all datasets. Figure 1 compares the in-distribution WGA (calculated at $\log \alpha^{te} = -0.6$) (solid lines) and the OOD WGA (at $\log \alpha^{te} = 0.6$) (dashed lines) throughout the learning process in **Source Datasets**.

Neural networks exploit extraneous artifacts for prediction. The generalization gap between in-distribution and OOD settings throughout the learning progress suggests that the predictor f leverages features in X that are independent of Y but correlated with Z to perform predictions. While this achieves high accuracy in-distribution, it results in limited generalizability to OOD settings (i.e., $f(x) \not\perp Z \mid Y$). This indicates a consistent shortcut learning pattern across **Source Datasets**, characterized by high in-distribution WGA at the beginning of training. Conversely, OOD WGA grows slowly or even degrades as training progresses, despite the fact that OOD settings share identical marginal distributions $P(Y)$ and $P(Z)$ with the in-distribution settings.

Checkpoint selection does not confer robustness to provenance shift. Model selection (selecting an optimal checkpoint during the training process) is a critical component of the learning pipeline (Gulrajani and Lopez-Paz, 2020). Oracle selection—which assumes access to the test distribution—often yields over-optimistic results and is generally not considered a valid benchmarking methodology for real-world deployment. Figure 1 shows WGA (y-axis) as training proceeds (x-axis). In this context, oracle selection would involve picking the training step along the x-axis that corresponds to the best WGA. As can be seen in Figure 1, our results highlight that even with oracle selection, ERM fails to reach a satisfactory solution under provenance shift. The stagnant OOD WGA curves in this figure show that there is no point in training at which ERM performance in the OOD setting approaches its in-distribution performance. This suggests that learning under provenance shift requires more than just better checkpointing; it necessitates a deliberate "deconfounding" process to prevent neural networks from using confounders as indicators of the label.

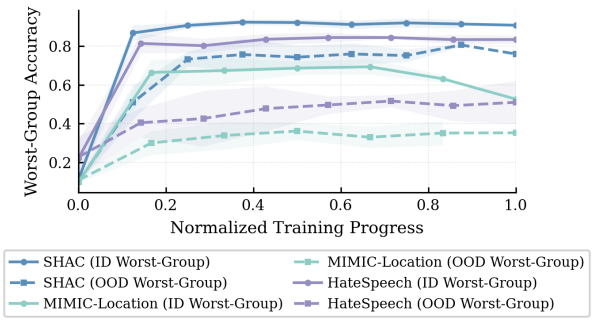

Figure 1: In-distribution (solid) and OOD (dashed) WGA for ERM in **Source Datasets**. Y-axis: Worst Group Accuracy (WGA). X-axis: normalized progress of training. The gap between the dashed and solid lines shows that models make inaccurate predictions in the OOD setting throughout their training process.

5.2. Predicting OOD Performance

Estimating and understanding the predictor f 's performance under provenance shift is a problem of interest for practitioners. The "Accuracy-on-the-line" phenomenon, which has been previously documented, describes the typical observation of a strong linear correlation between a model's in-distribution and OOD accuracy (Miller et al., 2021). While this is a useful heuristic for model selection (i.e., models with higher in-distribution accuracy are likely to have better OOD accuracy), this phenomenon is not consistent in some OOD benchmarks (Baek et al., 2022), and has not been explored in the context of provenance-specific label distribution shifts. In this section, we examine

1. Whether in-distribution WGA is a reliable indicator of OOD WGA under provenance shift?

2. Whether α^{te} can serve as a robust alternative OOD performance indicator?

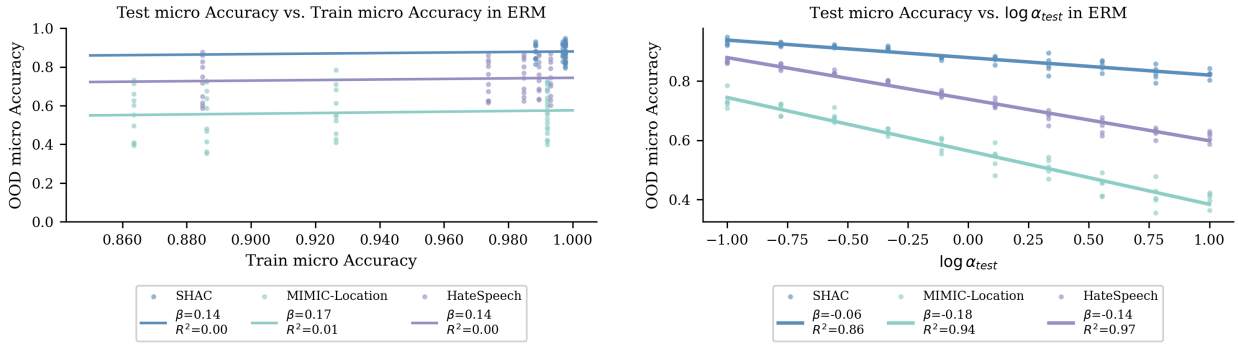

Figure 2: WGA is not "on the line" with respect to ID performance, but exhibits a strong linear relationship with the shift parameter α in **Source Datasets**.

Using a fixed predictor f per random seed, we calculated the OOD WGA across a suite of test distributions where $\log \alpha^{te}$ varies linearly from -1 to 1, while models are trained with $\log \alpha^{tr} = -0.6$. Results are shown in Figure 2, which plots the OOD WGA (y-axis) against the training WGA (left panel) and α (right panel). Lines have been fit to results for each model across multiple stochastically-initiated runs of the experiment, each illustrated with individual markers.

In-distribution WGA can underspecify OOD performance under provenance shift, while α is a strong OOD performance indicator. Similarly to Koh et al. (2021), we did not observe a strong correlation between in-distribution and OOD WGA across five runs with different random seeds (Figure 2). Instead, OOD performance varied significantly per run, as evidenced by the vertical scattering of OOD WGA relative to train WGA. This indicates that given a train WGA, the OOD WGA may not be predictable using in-distribution WGA, and improving WGA in distribution may not directly lead to improved performance in OOD settings, which is consistent with results reported in (Baek et al., 2022). Conversely, we found that α^{te} is strongly correlated with OOD test WGA across **Source Datasets**, yielding high R^2 values of 0.86, 0.94, and 0.97. Intuitively, since the predictor f was trained on a distribution with $\alpha^{tr} = -0.6$, test distributions with an α^{te} closer to α^{tr} generally exhibits higher performance. These results suggest that practitioners can estimate deployment robustness by performing a stress test to derive the linear relationship (coefficient and intercept) between OOD WGA and α.

5.3. Benchmarking Existing Algorithms

Invariant learning has garnered significant attention over the past decade; however, its efficacy under provenance shift remains largely unexplored. In this section, we address two primary questions:

1. Can existing algorithms successfully account for an observed binary confounder to learn a counterfactually invariant predictor $f(X) \perp Z \mid Y$ in OOD settings?

2. Do invariant learning algorithms remain robust under the provenance shift stress tests provided by DeconDTN-Toolkit?

To pursue answers to these questions, we evaluate a suite of invariant learning algorithms across **Source Datasets** in our toolkit. Following the same protocol as our ERM experiments, we perform a random search over the joint hyperparameter space for each algorithm and dataset. The optimal hyperparameter configurations are detailed in Table 6 and Table 7 for **Source Datasets**. The full evaluation results are detailed in Appendix G.

Training distribution manipulation was the most effective strategy to mitigate confounding under provenance shift. Our results indicate that Upsampling and Downsampling consistently improve OOD WGA across **Source Datasets**, often outperforming more complex invariant learning algorithms when evaluated in a consistent setting (Table 3). This aligns with previous findings by Idrissi et al. (2022a) and Hendrycks et al. (2020). Contrary to the prevailing expectation of improvement

Table 3: OOD WGA across algorithms in **Source Datasets**.

Algorithm	SHAC	MIMIC-Location	HateSpeech	Avg
ERM	0.83 ± 0.03	0.37 ± 0.06	0.51 ± 0.05	0.57
UpSampling	0.87 ± 0.02	0.47 ± 0.04	0.58 ± 0.04	0.64
DownSampling	0.87 ± 0.02	0.47 ± 0.05	0.61 ± 0.05	0.65
CAD	0.85 ± 0.02	0.38 ± 0.03	0.49 ± 0.04	0.57
CDANN	0.87 ± 0.01	0.37 ± 0.02	0.45 ± 0.03	0.56
CORAL	0.85 ± 0.05	0.40 ± 0.04	0.49 ± 0.04	0.58
DANN	0.85 ± 0.03	0.36 ± 0.03	0.52 ± 0.02	0.57
DFR	0.85 ± 0.02	0.35 ± 0.09	0.58 ± 0.04	0.60
DualFilter	0.82 ± 0.06	0.40 ± 0.05	0.51 ± 0.04	0.58
Fish	0.84 ± 0.03	0.36 ± 0.04	0.48 ± 0.04	0.56
GroupDRO	0.84 ± 0.03	0.36 ± 0.09	0.50 ± 0.05	0.57
IRM	0.85 ± 0.02	0.40 ± 0.05	0.51 ± 0.03	0.59
JTT	0.87 ± 0.01	0.38 ± 0.04	0.49 ± 0.06	0.58
LISA	0.87 ± 0.04	0.42 ± 0.05	0.56 ± 0.03	0.62
LfF	0.77 ± 0.08	0.45 ± 0.04	0.49 ± 0.06	0.57
MMD	0.85 ± 0.02	0.38 ± 0.04	0.51 ± 0.04	0.58
MTL	0.83 ± 0.04	0.37 ± 0.06	0.47 ± 0.05	0.56
Mixup	0.84 ± 0.03	0.37 ± 0.04	0.52 ± 0.03	0.58

as data scales, these results also show that discarding training samples (**Downsampling**) can be more effective than **Upsampling**, echoing the observations of Sagawa et al. (2020). Furthermore, algorithms incorporating balanced-sampling strategies—such as DFR and LISA—consistently outperform those without them. These observations suggest that while the underlying confounding structure $X \leftarrow Z \rightarrow Y$ persists, explicitly removing the correlation between Z and Y in the training distribution is a reliable path to robustness.

Gains from invariant learning algorithms are dataset-dependent and marginal compared to ERM. Overall, the evaluated algorithms achieve limited robustness to provenance shift across **Source Datasets**. While we observe some benefits from invariant learning methods (bottom panel of Table 3), these improvements are inconsistent across **Source Datasets** and often marginal relative to baseline ERM. This trend mirrors findings in domain generalization (Gulrajani and Lopez-Paz, 2020) and subpopulation shift (Yang et al., 2023). Consequently, developing methods to learn invariant features from joint-imbalanced distributions remains a critical open research direction under provenance shift.

5.4. Generalization to Attribute Datasets

In this section, we examine how the above findings generalize to other attributes in both clinical and non-clinical contexts under "provenance" shift. We repeat all experiments on **Attribute Datasets** using the identical pipeline to **Source Datasets**. Overall, the characteristics of provenance shift empirically generalize to other attributes (including sex, race, and genre) in both clinical and non-clinical context. We find that the generalization gap between in-distribution and OOD settings throughout the learning progress persisted, and oracle selection does not confer robustness to provenance shift in **Attribute Datasets** (Figure 3). In addition, in-distribution WGA can underspecify OOD performance under provenance shift, while α is a strong OOD performance indicator (Figure 4). Upsampling and Downsampling consistently improve OOD WGA on **Attribute Datasets**, often outperforming more complex invariant learning methods under a controlled evaluation setting. LISA achieves comparable WGA to subsampling on **Attribute Datasets**, while other methods yield only inconsistent and marginal gains over the ERM baseline (Table 10).

6. Discussion

We study the problem of provenance shift and introduce DeconDTN-Toolkit specializing in the problem setting. Our comprehensive empirical study using DeconDTN-Toolkit reveals that:

1. ERM consistently exploits provenance-related artifacts, leading to significant performance degradation in OOD settings.
2. Oracle selection fails to confer robustness under provenance shift. While in-distribution WGA is an unreliable predictor of OOD success, the shift parameter α offers a strong linear indicator for performance estimation.
3. Simple training distribution manipulations outperform complex invariant learning algorithms, highlighting the ongoing need for effective deconfounding methods.

We provide several practical recommendations under provenance shift and discuss several challenge for future research below:

Recommendation 1: Analyze the causal graph and tackle confounders at the study design phase. Constructing Directed Acyclic Graphs (DAGs) and mitigating confounding effects are standard practices in fields like epidemiology, yet research on deliberate modeling of confounders remains relatively limited within clinical NLP—particularly during data collection and ad-hoc modeling. Consequently, clinical NLP datasets that explicitly account for confounders are scarce, which impedes research on confounder adjustment in this context. Our work demonstrates the failure of ERM under provenance shift, emphasizes that collecting and accounting for confounders is critical for robustness in real-world deployment. Given the current lack of effective post-hoc solutions for unobserved confounders, we advocate that practitioners analyze the causal graph during the study design phase, prior to data collection. This practice can reduce the risk of collection bias, which induce spurious correlations between the outcome of interest and unobserved confounders.

Recommendation 2: Stress test the classifier under provenance shift before online deployment. Our study indicates the train-time metrics including WGA may not be strong predictors of OOD performance in a provenance shift setting. Consequently, we recommend that practitioners move beyond static in-distribution evaluations and adopt a provenance-aware stress testing protocol before deployment under the scenario of online learning, in which practitioners need to decide whether to retrain the model with the incoming online labeled samples. Practically, practitioners can synthetically vary the provenance shift parameter (α^{te}) to map the model's performance decay curve using DeconDTN-Toolkit during the ad-hoc testing phase. Deriving predictive coefficients can also allow for the estimation of performance in real-time by simply monitoring the α of incoming data. For example, in scenarios where online sample sizes are too small for traditional subsampling, stress testing serves as a proxy for empirical counterfactual invariance. If a model exhibits a steep performance drop as α^{te} deviates from α^{tr}, it indicates a high sensitivity to observed confounders that ID metrics would otherwise fail to flag. In this case, we would recommend exploring algorithms that are robust provenance shift, such as Backdoor Adjustment (Landeiro and Culotta, 2016, 2018; Ding et al., 2024) and DAPPER (Ding et al., 2025).

Challenge 1: Long-tailed categorical confounders demand a large sample size to remove spurious correlations in observational data. We study a simplified scenario where there is only one observed binary confounder. In contrast, consider a categorical confounder D with a set of categories $D = \{d_1, d_2, \ldots, d_k\}$, where the number of categories k is large (e.g., hundreds of rare comorbidities or specific medication types). In clinical datasets, the distribution of D is often highly skewed; while a few categories d_i are frequent, the majority reside in the "long tail" with minimal representation. With a such a "curse of rarity" in categorical confounders, removing spurious correlations between the outcome Y and the confounder D through subsampling becomes exponentially more difficult and the effective sample size is constrained by the rarest category necessary for the analysis.

Challenge 2: Removing confounding effects of continuous variables is still a open question. The current landscape of deconfounding research focuses primarily on categorical confounders and outcomes. However, real-world confounders—such as pollution levels, smoking intensity, or clinical care indices—are frequently continuous. When dealing with such continuous confounders, some of the aforementioned methods do not admit a natural extension. There is a critical need for differentiable deconfounding objectives that can integrate continu-

ous variable adjustments directly into neural network loss functions. Furthermore, the field lacks robust benchmarks that move beyond binary proxies to incorporate high-fidelity, continuous environmental and physiological metadata for deconfounding research.

7. Limitations

First, we acknowledge that achieving and evaluating counterfactual invariance are something of an ideal. With text categorization in particular it may not be possible to separate the features that indicate Y from those associated with Z and quantify their separation using observational data. Empirically, we observe geometric gaps between provenances in the representation space, which mirrors phenomenons in multimodal research (Liang et al., 2022) and can not be trivially eliminated by invariant learning algorithms such as MMD. Nonetheless, to the extent it is attainable, models with counterfactual invariance should be robust to provenance shift. In future work we will incorporate additional methods motivated by task arithmetic, which were designed to address confounding adjustment directly (Ding et al., 2025), and quantify the extent of counterfactual invariance using geometric separations in the representation space.

Second, we restrict our analysis to binary confounders Z and labels Y. Although this serves as an essential starting point, real-world causal structures involve multi-class or continuous variables; extending our framework to these richer settings is a key direction for future work. In addition, we only explored one direction of shift with a fixed α^{tr}, suggesting the need to explore shift directions for completeness. The shift direction is fully reversed when the sign of α change, which indicates positive samples comes from one to another. Ding et al. (2025) showed that performance patterns may vary with the shift direction. On account of this possibility, we advise simulating shifts in both directions for a complete characterization of robustness. Despite the varied relationships between shift directions and performance, we expect the linear relationship between α and OOD performance holds when the shift direction is fully reversed unless data from one source are more difficult to classify. This is consistent with prior evidence: Landeiro and Culotta (2016) observed a symmetric performance degradation when the shift direction is fully reversed in the general domain, as have Ding et al. (2025) with clinical data.

Third, our empirical findings are constrained by our controlled experimental design. To isolate the effects of provenance shift, we assume fixed marginals $P^{tr}(Y) = P^{te}(Y)$ and $P^{tr}(Z) = P^{te}(Z)$ with uniform distributions, which are necessary isolation for a first characterization of the problem. Furthermore, our observations are "sparse", focusing primarily on the association between $P(Y = 0)$ and $P(Z = 1)$; these results may not generalize to other associations or high-dimensional settings. As noted by Gulrajani and Lopez-Paz (2020), such negative claims are inherently restricted to the specific settings tested.

Finally, we adopt causal assumptions similar to Veitch et al. (2021), requiring that the invariant component $X_Z^{\perp}$ is statistically independent of Z. In practice, a common cause of $X_Z^{\perp}$ and Z could introduce dependencies, potentially leading to an over-conservative estimate of $\hat{X_Z^{\perp}}$ and the loss of predictive information. And a common cause of $X_Z^{\perp}$ and Y could introduce dependencies, potentially leading to an over-optimistic estimate of $\hat{X_Z^{\perp}}$ and the entanglement of provenance information. In addition, unobserved confounders U commonly pose a significant challenge in medical research. Here we discuss the scenario when $Y \leftarrow U \rightarrow Z$. In the anti-causal causal graph when $X_Z^{\perp} \leftarrow Y \leftarrow U \rightarrow Z$, the learning objective $f(X) \perp Z \mid Y$ is optimal and risk-invariant, while in the causal graph when $X_Z^{\perp} \rightarrow Y \leftarrow U \rightarrow Z$, the learning objective should be $f(X) \perp Z$ (Schrouff et al., 2024). The consideration of U falls outside the scope of the current work ($Y(z) \neq Y(z')$), but it represents an important direction for future work.

8. Conclusion

We establish a theoretical link between provenance shift, counterfactual invariance, and invariant learning, deriving a learning objective to achieve robustness under provenance shift. To facilitate empirical study, we introduce DeconDTN-Toolkit, a specialized and systematic suite to evaluate and enhance the robustness of text classification models to provenance shift. Our work provides both the theoretical grounding and the practical tools necessary to characterize the problem of confounding by provenance.

References

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. In *Advances in Neural Information Processing Systems*, volume 34, pages 3438–3450. Curran Associates, Inc., 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization, March 2020.

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, December 2022.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, February 2013. ISSN 0025-1909. doi: 10.1287/mnsc.1120.1641.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

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*, pages 491–500, 2019.

Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does Throwing Away Data Improve Worst-Group Error? In *Proceedings of the 40th International Conference on Machine Learning*, pages 4144–4188. PMLR, July 2023.

Annie S. Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. Project and Probe: Sample-Efficient Adaptation by Interpolating Orthogonal Features. In *International Conference on Learning Representations*, October 2023.

Benjamin Cohen-Wang, Joshua Vendrow, and Aleksander Madry. Ask Your Distribution Shift if Pre-Training is Right for You. *Transactions on Machine Learning Research*, June 2024. ISSN 2835-8856.

Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine learning for healthcare conference*, pages 110–127. PMLR, 2023.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Xiruo Ding, Zhecheng Sheng, Brian Hur, Feng Chen, Serguei VS Pakhomov, and Trevor Cohen. Enhancing robustness of foundation model representations under provenance-related distribution shifts. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

Xiruo Ding, Zhecheng Sheng, Meliha Yetişgen, Serguei Pakhomov, and Trevor Cohen. Backdoor Adjustment of Confounding by Provenance for Robust Text Classification of Multi-institutional Clinical Notes. *AMIA Annual Symposium Proceedings*, 2023:923–932, January 2024. ISSN 1942-597X.

Xiruo Ding, Zhecheng Sheng, Brian Hur, Justin Tauscher, Dror Ben-Zeev, Meliha Yetişgen, Serguei Pakhomov, and Trevor Cohen. Tailoring task arithmetic to address bias in models trained on multi-institutional datasets. *Journal of Biomedical Informatics*, 168:104858, August 2025. ISSN 1532-0464. doi: 10.1016/j.jbi.2025.104858.

Yana Dranker, He He, and Yonatan Belinkov. IRM—when it works and when it doesn' t: A test case of natural language inference. In *Advances in Neural Information Processing Systems*, volume 34, pages 18212–18224. Curran Associates, Inc., 2021.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable Domain Generalization via Quantile Risk Minimiza-

tion. *Advances in Neural Information Processing Systems*, 35:17340–17358, December 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. ISSN 1533-7928.

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, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, October 2020.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3: 642517, 2021.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained Transformers Improve Out-of-Distribution Robustness. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5149–5169, September 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3079209.

Kristen Howell, Megan Barnes, J Randall Curtis, Ruth A Engelberg, Robert Y Lee, William B Lober, James Sibley, and Trevor Cohen. Controlling for confounding variables: accounting for dataset bias in classifying patient-provider interactions. In *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, pages 271–282. Springer, 2020.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022a.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, June 2022b.

Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on artificial intelligence*, volume 56, pages 111–117, 2000.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: https://doi.org/10.1038/sdata.2016.35.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations, June 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664. PMLR, July 2021.

Michalis Korakakis, Andreas Vlachos, and Adrian Weller. Mitigating Shortcut Learning with InterpoLated Learning. In Wanxiang Che, Joyce Nabende,

Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9191–9206, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.450.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing Classifiers are Unreliable for Concept Removal and Detection. *Advances in Neural Information Processing Systems*, 35:17994–18008, December 2022.

Virgile Landeiro and Aron Culotta. Robust text classification in the presence of confounding bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Virgile Landeiro and Aron Culotta. Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*, 63:391–419, 2018.

Egoitz Laparra, Steven Bethard, and Timothy A Miller. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open*, 3(2): 146–150, 2020.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to Generalize: Meta-Learning for Domain Generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018a. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11596.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain Generalization with Adversarial Feature Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, June 2018b. doi: 10.1109/CVPR.2018.00566.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018c.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, October 2022.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3122–3130. PMLR, July 2018.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6781–6792. PMLR, July 2021.

Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? *arXiv preprint arXiv:2403.08632*, 2024a.

Zhuang Liu and Kaiming He. A Decade's Battle on Dataset Bias: Are We There Yet? In *The Thirteenth International Conference on Learning Representations*, October 2024b.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631, 2021.

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 *Proceedings of the 38th International Conference on Machine Learning*, pages 7721–7735. PMLR, July 2021.

Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. Explicit Tradeoffs between Adversarial and Natural Distributional Robustness. *Advances in Neural Information Processing Systems*, 35:38761–38774, December 2022.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. ISSN 0003-4851.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant Gradient Variances for Out-of-Distribution Generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18347–18377. PMLR, June 2022.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647.

Yangjun Ruan, Yann Dubois, and Chris J. Maddison. Optimal Representations for Covariate Shift. In *International Conference on Learning Representations*, October 2021.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, September 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8346–8356. PMLR, November 2020.

Jessica Schrouff, Alexis Bellot, Amal Rannen-Triki, Alan Malek, Isabela Albuquerque, Arthur Gretton, Alexander D'Amour, and Silvia Chiappa. Mind the graph when balancing data for fairness or robustness. *Advances in Neural Information Processing Systems*, 37:29913–29947, 2024.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y Chen. The data addition dilemma. *arXiv preprint arXiv:2408.04154*, 2024.

Zhecheng Sheng, Xiruo Ding, Brian Hur, Changye Li, Trevor Cohen, and Serguei V. S. Pakhomov. Mitigating Confounding in Speech-Based Dementia Detection through Weight Masking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10419–10434, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.514.

Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient Matching for Domain Generalization. In *International Conference on Learning Representations*, October 2021.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8. doi: 10.1007/978-3-319-49409-8_35.

Damien Teney, Jindong Wang, and Ehsan Abbasnejad. Selective Mixup Helps with Distribution Shifts, But Not (Only) because of Mixup. In *Proceedings of the 41st International Conference on Machine Learning*, pages 47948–47964. PMLR, July 2024.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, June 2011. doi: 10.1109/CVPR.2011.5995347.

Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual Invariance to Spurious Correlations in Text Classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208. Curran Associates, Inc., 2021.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics*

and the 11th international joint conference on natural language processing (volume 1: long papers), pages 1667–1682, 2021.

Yanshan Wang, Jordan Hilsman, Chenyu Li, Michele Morris, Paul M Heider, Sunyang Fu, Min Ji Kwak, Andrew Wen, Joseph R Applegate, Liwei Wang, et al. Development and validation of natural language processing algorithms in the national enact network. *Journal of Clinical and Translational Science*, 9(1):e199, 2025.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 1112–1122, 2018.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve Unsupervised Domain Adaptation with Mixup Training, January 2020.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is Hard: A Closer Look at Subpopulation Shift. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39584–39622. PMLR, July 2023.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving Out-of-Distribution Robustness via Selective Augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 25407–25437. PMLR, June 2022.

Kotaro Yoshida and Hiroki Naganuma. Towards Understanding Variants of Invariant Risk Minimization through the Lens of Calibration. *Transactions on Machine Learning Research*, April 2024. ISSN 2835-8856.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, February 2018.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and Combating Spurious Features under Distribution Shift. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12857–12867. PMLR, July 2021.

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, April 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3195549.

Appendix A. Proof

A.1. Proof of Lemma 2

Lemma 2 (Prediction Decomposition). *Let $X_Y^{\perp}$ be Y-invariant components of the input features, such that $X_Y^{\perp}(y) = X_Y^{\perp}(y')$. A discriminative model under the causal graph $\mathcal{G} : X \leftarrow Z \rightarrow Y$ can be decomposed into two components: the inference and the provenance mechanism.*

$$P(Y \mid X_Y^{\perp}) = \int \underbrace{P(Y \mid Z)}_{provenance} \underbrace{P(Z \mid X_Y^{\perp})}_{inference} dZ \quad (4)$$

Proof By the law of total probability, we marginalize over the Z:

$$P(Y \mid X_Y^{\perp}) = \int P(Y \mid Z, X_Y^{\perp}) P(Z \mid X_Y^{\perp})\, dZ$$

By definition, $X_Y^{\perp}$ is Y-invariant, implying $P(Y \mid Z, X_Y^{\perp}) = P(Y \mid Z)$. Therefore, the conditional probability simplifies to $P(Y \mid Z, X_Y^{\perp}) = P(Y \mid Z)$. Substituting this back into the integral yields:

$$P(Y \mid X_Y^{\perp}) = \int \underbrace{P(Y \mid Z)}_{\text{provenance}} \underbrace{P(Z \mid X_Y^{\perp})}_{\text{inference}} dZ$$

■

A.2. Proof of Proposition 4

Proposition 4 (Provenance Robustness). *If a predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$ satisfies counterfactual invariance such that $f(X(z)) = f(X(z'))$ for all $z, z' \in \mathcal{Z}$, then the predictor is robust to provenance shift. Specifically, the risk $\mathbb{E}[\mathcal{L}(f(X), Y)]$ remains constant under any intervention on the mechanism $P(Y|Z)$, provided $P(Y)$ remains invariant. This robustness holds under both:*

1. *Anti-causal settings ($Y \rightarrow X$);*
2. *Causal settings ($X \rightarrow Y$): Provided that $Y \perp X \mid \{X_Z^{\perp}, Z\}$ and the label satisfies counterfactual consistency, $Y(z) = Y(z')$ for all $z, z' \in \mathcal{Z}$.*

Proof We marginalize the probability of a prediction $\hat{y}$ over the joint distribution of Y and Z:

$$P(f(X) = \hat{y}) = \int_{\mathcal{Y}} \int_{\mathcal{Z}} P(f(X) = \hat{y} \mid Y, Z) P(Y, Z)\, dZ\, dY \quad (5)$$

$$= \int_{\mathcal{Y}} \int_{\mathcal{Z}} P(f(X) = \hat{y} \mid Y, Z) P(Y \mid Z) P(Z)\, dZ\, dY \quad (6)$$

Theorem 5 (Counterfactual Invariant Predictor) *(Restated from Thereom 3.2 in Veitch et al. (2021)) If f is a counterfactual invariant predictor,*

1. *Under the anti-causal graph, then $f(X) \perp Z \mid Y$.*
2. *Under the causal-direction graph, if Y and Z are not subject to selection (but possibly confounded), then $f(X) \perp Z$.*
3. *Under the causal-direction graph, if $Y \perp X \mid \{X_Z^{\perp}, Z\}$ and $Y(z) = Y(z')$ for all $z, z' \in \mathcal{Z}$, then $f(X) \perp Z \mid Y$.*

Applying the conditional independence property $f(X) \perp Z \mid Y$, which holds in the above two causal graphs, the term $P(f(X) = \hat{y} \mid Y, Z)$ simplifies to $P(f(X) = \hat{y} \mid Y)$:

$$P(f(X) = \hat{y}) = \int_{\mathcal{Y}} P(f(X) = \hat{y} \mid Y) \left(\int_{\mathcal{Z}} P(Y \mid Z) P(Z)\, dZ \right) dY \quad (7)$$

$$= \int_{\mathcal{Y}} P(f(X) = \hat{y} \mid Y) P(Y)\, dY \quad (8)$$

The final expression shows that the distribution of predictions (and consequently the risk) depends only on the model performance $P(\hat{Y} \mid Y)$ and the label prior $P(Y)$. Since the provenance mechanism $P(Y|Z)$ has been marginalized out, the predictor is robust to shifts in that mechanism. ■

Appendix B. Related Work

In this section, we provide an exhaustive literature review on possible solutions to provenance shift from different perspectives.

B.1. Distribution Alignment

Distribution alignment techniques align high-level statistics (e.g., mean) of distance measurement on features across provenances by minimizing their differences. For example, the Maximum Mean Discrepancy (MMD) measures and then minimizes the probability divergence between two distributions by mapping samples to a reproducing kernel Hilbert space (RKHS, e.g., using Gaussian kernels) and then deriving the difference of distribution across provenances (Li et al., 2018b). Similarly, Sun and Saenko (2016) proposed deep CORAL, which aligns the mean and the variance of the features instead of using parametric Gaussian kernels as in MMD. From a causal

perspective, Veitch et al. (2021) theoretically proves that domain alignment should consider the causal direction between the data and label. Specifically, when modeling causal direction data (i.e., the data causes the label), features from different provenances should be aligned like the classical practice. However, when modeling anti-casual direction data (i.e., the label causes the data), features from different provenances should be aligned *conditioned on the label*[9]

B.2. Adversarial Training

Compared to distribution alignment techniques, which make low-dimensional statistics indiscriminate to provenance, adversarial training techniques make the full feature space indiscriminate to provenance using a parametric provenance discriminator (i.e., predictor) based on neural networks. Adversarial training was initially proposed to train a generative model, which generates photorealistic images using random noise (Goodfellow et al., 2014). Specifically, adversarial training resembles a two-player game, which trains a *discriminator* to distinguish between real and the generated fake images by minimizing the real-fake classification loss, while rewarding the *generator* to fool the discriminator by maximizing the real-fake binary classification loss. With this training paradigm, the generator will finally generate photorealistic images that can fool a strong real-fake image discriminator. Ganin et al. (2016) extends this idea to domain adversarial training by learning domain-agnostic features that can confuse a domain discriminator. Specifically, Ganin et al. (2016) proposed the Domain Adversarial Neural Network to learn features discriminative for the main learning task but indiscriminate with respect to provenances. Li et al. (2018c) proposed Conditional DANN (CDANN) to adapt to concept shift by making the domain discriminator predict the permutation of provenance and label. In this case, the domain discriminator can't tell the features from different provenances but with the same label apart. Ruan et al. (2021) extends the domain adversarial training objective and proposes the Contrastive Adversarial Domain (CAD) objective by explicitly considering the discriminative capability of features. Specifically, the optimal features should remain discriminative for the learning task while distributions of features are indiscriminative to provenances in the training objective.

Similarly, adversarial training on word embedding makes the word embedding indiscriminative to the provenance in a post-hoc way. For example, Iterative Null-space Projection (INLP) iteratively removes an attribute from the word embedding space by projecting it to an attribute-agnostic subspace (Ravfogel et al., 2020). Specifically, it first trains a linear provenance discriminator and then projects the feature on the null-space of the provenance discriminator, which is expected to be non-discriminative of the provenance.

Though adversarial training is a strong method to make the feature provenance-agnostic, it can increase the reliance on spurious features instead of core features in a neural network with regulation (e.g. $\updownarrow_2$ regulation) (Moayeri et al., 2022). Besides, adversarial training can have an unintended consequence in reducing robustness to distribution shift, specifically when spurious correlations are changed in the test distribution. Kumar et al. (2022) argue that both adversarial training and word embedding editing methods can be counter-productive in real-world settings where the label is naturally correlated with the provenance, because they internally use an auxiliary (or probing) provenance classifier based on the features learnt by the main-task classifier. Specifically, the provenance classifier cannot be a reliable signal on whether the feature is causally derived from the provenance.

B.3. Invariant Learning

The Invariant Risk Minimization (IRM) objective is a variant of the ERM, which finds the features such that the optimal main-task classifier based on these features is **simultaneously** optimal for all provenances To overcome the challenging and bi-level optimization problem, Arjovsky et al. (2020) proposed a practical version, IRMv1, which penalizes the risk variation across provenances using one single scalable parameter. Dranker et al. (2021) conducts a case study in natural language inference and finds that learning complex features in place of spurious correlation proves to be difficult in the wild, leading to little incremental over ERM. In addition, the performance of IRM depends on the sample size, the prevalence of spurious correlation, and the strength of spurious correlation, therefore limiting IRM's advantage in the wild. Ahuja et al. (2021) adds infor-

9. Here, we use the MMD, denoted as MMD, for alignment as an example. Formally, if $X \to Y$, then the training objective is $\arg\min_X \text{MMD}(X)$. If $X \leftarrow Y$, then the training objective is $\arg\min_X \text{MMD}(X \mid Z)$.

mation bottleneck constraints (Tishby et al., 2000) to the IRM to mitigate failures when invariant features capture all the information about the label (e.g., the label is a deterministic function of the invariant feature). An empirical study shows that the Information Bottleneck-based IRM achieves consistent calibration across provenances and information compression techniques (e.g., information bottleneck) are potentially effective in achieving invariance (Yoshida and Naganuma, 2024).

B.4. Meta Learning

Unlike conventional machine learning approaches that use a *fixed* learning algorithm to solve the question from scratch, meta learning, known as learning-to-learning, aims to improve the learning algorithm itself using experience of multiple learning episodes (Hospedales et al., 2022). The motivation behind meta learning for distribution shift is that exposing a model to distribution shift can make it learn to adapt to the shifted domain. For example, the Meta-Learning Domain Generalization approach (MLDG) simulates distribution shift by synthetically perturbing the pseudo testing distribution in each training step (Li et al., 2018a). The meta-optimization objective is that the performance on the training distribution should also improve with the pseudo testing distribution simultaneously.

B.5. Gradient Matching

Gradients are signals that control learning orientation in neural network training. The idea of the gradient matching method under provenance shift is to align the provenance-specific gradient directions to derive provenance-invariant optimization paths and features. Shi et al. (2021) propose an Inter-Domain Gradient Matching (IDGM) objective that maximizes the inner product between gradients of different provenance. To overcome the computational cost of the second-order derivatives, Shi et al. (2021) propose Fish, a meta-learning optimization method based on simply first-order derivatives, to optimize the IDGM objective. Rame et al. (2022) extends the IDGM objective to the Fishr regularization, in which they match the provenance-specific gradient variances to match provenance-level risks and Hessians.

B.6. Distributionally Robust Optimization

Group Distributionally Robust Optimization (Group DRO) is a principle that minimizes the *worst-case* over potential test distributions training error instead of the *average* training error in classical ERM (Ben-Tal et al., 2013). Overparameterization refers to increasing model size beyond the point of zero training error, which can exacerbate spurious correlations when they are present in the training data (Sagawa et al., 2020). Sagawa* et al. (2019) study group DRO in the context of overparameterized neural networks, which achieve zero training error but don't generalize to the worst group at testing time. They found that strongly-regularized group DRO models without vanishing training signals have good worst-case performance. Besides, Sagawa* et al. (2019) developed a stochastic, online, and greedy algorithm for group DRO optimization that can scale to large models and datasets. Zhou et al. (2021) propose Group Conditional DRO (GC-DRO) to introduce group uncertainty in the case when pre-defined group information does not directly account for various spurious correlations. Specifically, GC-DRO introduces group-level and group-conditioned sample-level weights in the training process to generate a more flexible uncertainty set compared to group DRO, which treats each group as a unit and thus removes sample-level weights *within each group*. Inspired by the Group DRO, Eastwood et al. (2022) propose Quantile Risk Minimization (QRM) objective, which seek predictors that perform well with a specific probability α. In other words, while ERM seeks predictors that perform well on the *average-case* and Group DRO seeks predictors that perform well on the *worst-case*, QRM seeks *probabilistic* predictors that perform well with probability α by minimizing the α-quantile of the estimated risk distribution over training domains. Specifically, Eastwood et al. (2022) proposed the Empirical QRM (EQRM), which leverages kernel density estimation (KDE, Parzen (1962)) to estimate the cumulative distribution function (CDF) of the training risk and minimize the probability at the α-quantile of the CDF.

B.7. Double-Stage Training

Double-phase training strategies add a fine-tuning step after the standard ERM training. Just Train Twice (JTT) retrains the neural network using a reweighted dataset (second stage), in which samples misclassified at the end of a few steps of the stan-

dard ERM (first stage) are upweighted (Liu et al., 2021). Different from most mentioned methods, JTT does not require group information during training time. Intuitively, JTT improved the worst-group performance by focusing on samples from groups where standard ERM models perform poorly. Kirichenko et al. (2023) observed that ERM can learn core features even when spurious correlations are present and are much simpler than the core features in the training data. In addition, they noticed the final classification layer of the model highly weights the spurious features, resulting in poor predictions on the minority groups. Therefore, Kirichenko et al. (2023) proposed Deep Feature Reweighting (DFR), which leveraged a small set of data where the spurious correlation does not hold to retrain the classifier exclusively after the standard ERM training. Following DFR, Chen et al. (2023) projects the features in an orthogonal space and then trains the domain-specific classifiers using small data sets from the target domain. This approach is theoretically sample-efficient in training the domain-specific classifier with minimal distribution assumptions.

B.8. Re-Sampling and Re-Weighting

Re-sampling and re-weighting strategies manipulate the training distribution to prevent the model from learning from spurious correlation. Re-sampling simply down-samples and re-weighting simply upweights the sample according to the number of samples per class or group (Japkowicz, 2000). In other words, re-sample throws away data points until the classes or groups are balanced in size, followed by ERM on the down-sampled dataset. While discarding data opposes common wisdom in learning theory, where the expected error is inversely proportional to the sample size, Chaudhuri et al. (2023) theoretically shows that minor groups (i.e., the tail of the data distribution) play an important role in determining the worst-group-accuracy of a classifier that classifies samples with the largest possible gap between the labels. In addition, resampling outperforms ERM in worst-group-error when learning from imbalanced classes with tails and balanced classes but imbalanced groups. In the previous work, Gulrajani and Lopez-Paz (2020) implements a group-balanced ERM and found that it outperforms the state-of-the-art in terms of average performance across datasets, and the improvement of existing algorithms is too incremental compared to the group-balanced ERM. This finding is consolidated by Hendrycks et al. (2020), who found that if the samples of different datasets were unbiasedly drawn from the same distribution, the model should not discover any dataset-specific patterns. Similarly, Sagawa et al. (2020) finds that subsampling the majority group can empirically achieve low minority error in the overparameterized regime, but upweighting the minority group fails. Idrissi et al. (2022b) extends the work by considering both class-balance and group-balance and finds that re-sampling and re-weighting are competitive in common benchmarks in which classes or groups are imbalanced. Re-sampling and re-weighting are advanced and are recommended in the wild as they are faster to train and are hyper-parameter-free. Cohen-Wang et al. (2024) study pre-trained models and find that fine-tuning on a small but balanced dataset can result in significantly more robust models than fine-tuning on a large but imbalanced dataset.

B.9. Domain Interpolation

Mixup linearly combines, or *interpolates*, two random samples and applies the same interpolation strategy on the corresponding labels to create neighborhood (or *vicinity*) samples, thereby overcoming the limitations of dataset-dependent and intra-label data augmentation (Zhang et al., 2018). Yan et al. (2020) extended this idea to inter-domain Mixup, which linearly combines two random samples from different domains by explicitly considering the sample provenance. In a case study of natural language understanding tasks using inter-domain Mixup, this method was empirically demonstrated to improve the generalizability of minor groups across encoder, encoder-decoder, and decoder-only architectures consistently (Korakakis et al., 2025). Following inter-domain Mixup, Learning Invariant Predictors with Selective Augmentation (LISA) combines inter-label Mixup and inter-domain Mixup by randomly selecting one strategy for augmentation to introduce cross-label augmentation to inter-domain Mixup (Yao et al., 2022). Although empirical results of LISA have shown remarkable improvements in several benchmarks, Teney et al. (2024) pointed out that the performance gain of LISA might originate from implicitly balancing the training distribution instead of interpolated mixing. For example, intra-label Mixup implicitly resamples the training data to a class-uniform distribution, which is a strategy in tackling label shift.

Appendix C. Algorithms

In this section, we denote $\mathcal{X}$ as the input space, $\mathcal{Y}$ as the label space, and $\mathcal{H}$ as the representation space. Let $\Phi : \mathcal{X} \to \mathcal{H}$ represent the featurizer and $w : \mathcal{H} \to \mathcal{Y}$ the classifier, such that $f = w \circ \Phi$ defines the composite predictor. The loss function is denoted by $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

ERM ERM serves as the standard baseline for supervised learning. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, the learning objective is to find a predictor $f \in \mathcal{F}$ that minimizes the empirical risk:

$$\min_f \mathbb{E}_{(x,y)\sim P^{tr}}[\mathcal{L}(f(x), y)] \tag{9}$$

UpSampling and DownSampling (Japkowicz, 2000) UpSampling and DownSampling share the ERM objective but modify the training distribution P^{tr} to produce a balanced joint distribution $\hat{P}(X, Y, Z)$. Specifically, the target distribution is adjusted such that:

$$\hat{P}(X, Y, Z) = P(X, Y, Z)\frac{P(Y)P(Z)}{P(Y, Z)} \tag{10}$$

This adjustment aims to decouple the dependence between labels Y and confounder Z. A detailed discussion of specific sampling implementations is provided in Appendix B.8.

DANN and CDANN (Ganin et al., 2016; Li et al., 2018c) DANN introduce a domain discriminator $f_{disc} : \mathcal{H} \to \mathcal{Z}$ to ensure the representations are uninformative of the domain $z \in \mathcal{Z}$. The objective is:

$$\begin{aligned}\min_{w,\Phi} \max_{f_{disc}} &\mathbb{E}_{(x,y)\sim P^{tr}}[\mathcal{L}(w(\Phi(x)), y)] \\ &- \alpha\mathbb{E}_{(x,z)\sim P^{tr}}[\mathcal{L}_{disc}(f_{disc}(\Phi(x)), z)]\end{aligned} \tag{11}$$

where $\mathcal{L}_{disc}$ is the binary cross-entropy loss for domain classification and α is a trade-off hyperparameter.

CDANN conditions the domain discriminator by adding a label-specific embedding to the provenance discriminator input. Denote $\mathbf{e}_y$ as the embedding vector associated with label y. The objective of CDANN is:

$$\begin{aligned}\min_{w,\Phi} \max_{f_{disc}} &\mathbb{E}_{(x,y)\sim P^{tr}}[\mathcal{L}(w(\Phi(x)), y)] \\ &- \alpha\mathbb{E}_{(x,y,z)\sim P^{tr}}[\mathcal{L}_{disc}(f_{disc}(\Phi(x), \mathbf{e}_y), z)]\end{aligned} \tag{12}$$

A discussion of adversarial techniques is provided in Appendix B.2.

MMD and CORAL (Sun and Saenko, 2016; Li et al., 2018b) CORAL and MMD penalize the distribution distances across provenances. CORAL minimizes the distance between the second-order statistics of the representations across provenances. Let C be the covariance matrices of the source and target features in $\mathcal{H}$. The CORAL penalty is defined using the squared Frobenius norm:

$$\mathcal{L}_{\text{CORAL}} = \sum_{i,j\in\mathcal{Z}} \frac{1}{4d^2}\|C_i - C_j\|_F^2 \tag{13}$$

where d is the dimension of the representations.

MMD is a non-parametric metric that measures the discrepancy between two distributions P_s and P_t by comparing their mean embeddings in an RKHS $\mathcal{F}$ associated with a kernel $k(\cdot, \cdot)$. The squared MMD distance is:

$$\mathcal{L}_{MMD} = \sum_{i,j\in\mathcal{Z}} \left\|\mathbb{E}_{x\sim P_i}[\phi(x)] - \mathbb{E}_{u\sim P_j}[\phi(u)]\right\|_{\mathcal{F}}^2 \tag{14}$$

A discussion of domain alignment techniques is provided in Appendix B.1.

CAD (Ruan et al., 2021) The CAD learns to keep the representation H discriminative for the learning task and maintain the support of its marginal distribution invariant to shifts. Denote A is a random variable sampled conditionally from X. CAD instead maximizes the mutual information $I[A; \Phi(X)]$ based on InfoNCE (Oord et al., 2018) as an alternative learning objective. In addition, CAD introduces a domain bottleneck $I[\Phi(X); Z]$, which enforces support match using a KL divergence.

Mixup and LISA (Zhang et al., 2018; Yao et al., 2022) Mixup linearly interpolates one sample (x_i, y_i) with another random sample (x_j, y_j) from the training data:

$$\min_f \mathbb{E}_{(x,y)\sim P^{tr}} \mathcal{L}(f(\hat{x}), \hat{y}) \tag{15}$$

where $\hat{x} = \lambda x_i + (1 - \lambda)x_j$, $\hat{y} = \lambda\hat{y}_i + (1 - \lambda)\hat{x}_j$, and $\lambda \sim \text{Beta}(\alpha, \alpha)$, for $\alpha \in (0, \infty)$.

LISA explicitly consider label and provenance in the Mixup process by mixing samples (1) with the same provenance but different labels (i.e., intra-domain $y_i \neq y_j$ and $z_i = z_j$) and (2) with the same provenance but different labels (i.e., intra-label $y_i = y_j$ and $z_i \neq z_j$). A discussion of domain interpolation techniques is provided in Appendix B.9.

IRM (Arjovsky et al., 2020) IRM penalizes feature distributions that have different optimal linear classifiers for each domain. The IRM objective is

$$\begin{aligned} &\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{z\in\mathcal{Z}} \mathbb{E}_{(x,y)\sim P_z}[\mathcal{L}(w\circ\Phi(x),y)] \\ &\text{s.t.} \quad \min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} \mathbb{E}_{(x,y)\sim P_z}[\mathcal{L}(\bar{w}\circ\Phi(x),y)], \forall z\in\mathcal{Z}. \end{aligned} \tag{16}$$

A discussion of invariant learning techniques is provided in Appendix B.3.

GroupDRO (Sagawa* et al., 2019) GroupDRO uses distributionally robust optimization to explicitly minimize the loss on the worst-case domain during training:

$$\min_f \max_{z\in\mathcal{Z}} \mathbb{E}_{(x,y)\sim P_z^{tr}}[\mathcal{L}(f(x),y)] \tag{17}$$

In practice, this is implemented using an online estimation of group weights q, where q is updated via exponentiated gradients to shift the model's focus toward groups with higher training error. A discussion of distributionally robust techniques is provided in Appendix B.6.

Fish (Shi et al., 2021) Fish proposed an inter-domain gradient matching objective to align the gradient direction across provenances. To reduce the computing complexity of second-order derivatives, Fish use a first-order algorithm for approximation. Similar to meta learning, Fish updates a clone $\hat{f}$ using data per provenance and then update using a weighted difference bwteen the clone model $\hat{f}$ and the modek before update f. A discussion of gradient matching techniques is provided in Appendix B.5.

MTL (Blanchard et al., 2021) MTL augments the representation with the marginal distribution of feature vectors. Specifically, MTL maintains a domain embedding $\mathbf{e}$, which is the empirical mean of the representations in that domain:

$$\mathbf{e}_z = \mathbb{E}_{x\sim P_z^{tr}}[\Phi(x)] \tag{18}$$

In practice, $\mathbf{e}_z$ is updated via an exponential moving average (EMA) during training. The MTL learning objective is based on the concatenation of the pointwise feature with the domain embedding:

$$\min_f \mathbb{E}_{(x,y)\sim P_z^{tr}}[\mathcal{L}(w\left(\|\Phi(x);\mathbf{e}_z\|_2\right),y)] \tag{19}$$

where $[\cdot;\cdot]$ denotes concatenation. In inference, $\mathbf{e}_z$ degrades to the mean of $\Phi(x)$. This allows the classifier to adapt its decision boundary based on the global statistics of the current marginal distribution.

LfF (Nam et al., 2020) LfF simultaneously trains a biased classifier f_B and a debiased classifier f_D. The biased model is optimized by $\mathcal{L}_{\text{GCE}}$ to amplify bias:

$$\mathcal{L}_{\text{GCE}} = \mathbb{E}_{(x,y)\sim P^{tr}} \frac{1-P\left(f(x)=y\right)^q}{q} \tag{20}$$

where $q\in(0,1]$ is a hyperparameter that controls the amplification degree. The debiased model is optimized by

$$\mathcal{L}_{\text{LfF}} = \mathbb{E}_{(x,y)\sim P^{tr}} \frac{\mathcal{L}(f_B(x),y)}{\mathcal{L}(f_B(x),y)+\mathcal{L}(f_D(x),y)} \mathcal{L}(f_D(x),y) \tag{21}$$

JTT (Liu et al., 2021) JTT first curates an error set E of training samples that the ERM model f_{ERM} misclassifies:

$$E = \{(x_i,y_i) \text{ s.t. } f_{\text{ERM}}(x_i)\neq y_i\} \tag{22}$$

Then JTT upweight the samples in E:

$$\min_f \lambda \sum_{(x,y)\in E} \mathcal{L}(f(x),y) + \sum_{(x,y)\notin E} \mathcal{L}(f(x),y) \tag{23}$$

where λ is a hyperparameter.

DFR (Kirichenko et al., 2023) DFR adapts the standardized ERM training procedure. It then freezes the featurizer Φ and and retrains the classifier w with l_2 penalization using a subset from $\mathcal{D}^{tr}$. The subset is sampled to follow a joint balanced distribution $\hat{P}^{tr}$.

BackDoor (Ding et al., 2024) BackDoor includes a provenance embedding in the model to conduct backdoor adjustment. The prediction can be decomposed as

$$P(Y\mid X) = \sum_{z\in\mathcal{Z}} P(Y\mid X,z)P(z) \tag{24}$$

Backdoor uses $P^{tr}(Z)$ as $P(Z)$ in inference.

DualFilter (Sheng et al., 2025) DualFilter first trains a task classifier from the pretrained weights θ_0, obtaining the accumulative weight changes Δ_{task} across the network and finetuned weights $\hat{\theta}$; and then it trains a provenance classifier g from the same pretrained checkpoint with accumulative weight change Δ_{prov}. For both Δ, pick the top k most changed weights locations for set operation and masking.

$$\begin{aligned} &\text{Let } M = \Delta_{task,k} \odot \Delta_{prov,k} \\ &f_{\theta'}: \theta'_i \leftarrow \hat{\theta}_i = 0 \quad \forall i\in M, \end{aligned} \tag{25}$$

where $\odot$ is an arbitrary set operation and $f_{\theta'}$ is the masked task model ready for inference.

Appendix D. Datasets

SHAC (Lybarger et al., 2021) The Social History Annotation Corpus (SHAC) consists of clinical notes from two institutions: the University of Washington Medical Center and MIMIC-III. The primary task is to identify information regarding substance use from clinical text. In our experiments, we define the prediction label as the presence of drug abuse and utilize the data source (institution) as the provenance.

MIMIC-Location (Johnson et al., 2016) **MIMIC-Location** contains clinical notes recorded during the first 48 hours of a hospital stay, sourced from the MIMIC-III database. We define the prediction target as in-hospital mortality and use admission location (Emergency Room Admission vs. Physician Referral / Normal Delivery) as the provenance.

HateSpeech (Vidgen et al., 2021; De Gibert et al., 2018) We utilize a hate speech detection dataset curated by Ding et al. (2025), which aggregates samples from two distinct sources: synthetically generated text and posts from a white supremacist forum. We treat toxicity detection as the prediction task and use the data source as the provenance

Civilcomments (Borkan et al., 2019) **Civilcomments** is a comment collection from online articles. identities that are mentioned in the comment We follow the same preprocessing procedure in WILDS (Koh et al., 2021). We use the mention of demographics as a proxy to stereotyping on certain attribute which is a common cause of the toxicity and the comment. We select the mention of black as the provenance.

MultiNLI (Williams et al., 2018) **MultiNLI** is a Natural Language Inference (NLI) dataset designed to predict the logical relationship (entailment, neutral, or contradiction) between a premise and a hypothesis. Using the training subset, we binarize the task to predict non-entailment (grouping neutral and contradiction) and define the provenance based on whether the genre is "fiction."

MIMIC-SubpopBench (Yang et al., 2023) **MIMIC-SubpopBench** contains clinical notes recorded during the first 48 hours of a hospital stay, sourced from the MIMIC-III database. We follow the preprocessing pipeline from SubpopBench We define the prediction target as in-hospital mortality and use patient sex as the provenance.

Appendix E. Experiments

E.1. Datasets

We subsampled the datasets to introduce spurious correlation in the training distribution. We use the sampling parameters $\log \alpha^{tr} = \log \alpha^{val} = -0.6$, $-1 \leq \log \alpha^{te} \leq 1$, $|\mathcal{D}^{tr}| : |\mathcal{D}^{val}| : |\mathcal{D}^{te}| = 6 : 2 : 2$. With these configurations, we derive 2,162 samples from **SHAC**, 1,697 samples from **MIMIC-Location**, 6,996 samples from **HateSpeech**, 3,681 samples from **MIMIC-SubpopBench**, 8,207 samples in **Civilcomments**, and 62,120 samples from **MultiNLI**. We alter the random seed in the subsampling process to make the subsets are seed-dependent and algorithm-independent (i.e., algorithms are trained and compared using the same subset).

E.2. Hyperparameters

We list the joint distribution of hyperparameters per algorithm for random search in Table 4.

E.3. Algorithm

We conduct our evaluation using all algorithms available in the toolkit, with the notable exception of Backdoor Adjustment (Landeiro and Culotta, 2016, 2018), an algorithm deliberately developed to address confounding shift that has demonstrated utility in addressing confounding by provenance (Ding et al., 2024). We do not include this algorithm in the current evaluation because it is not intended for use in the context of end-to-end fine-tuned deep learning models for NLP, and on account of this underperforms relative to the algorithms under consideration here. OOD WGA of BackDoor on each dataset was 0.38, 0.53, 0.39, 0.33, 0.52, as compared to ERM (0.51 ± 0.05, 0.51 ± 0.05, 0.28 ± 0.02, 0.36 ± 0.05, 0.83 ± 0.03) on the **Civilcomments**, **HateSpeech**, **MIMIC-SubpopBench**, **MultiNLI**, and **SHAC** datasets respectively, using the default hyperparameter.

E.4. Training Details

We fine-tune a BERT model (`bert-base-uncased`) for all experimental settings (Devlin et al., 2019). Following the protocol in Yang et al. (2023), we employ early stopping based on the validation WGA with patience of three checkpoints. This strategy is applied to both stages of two-stage algorithms, including JTT, DFR, and DualFilter.

We train 3,000 steps on MultiNLI, 1,000 steps on **MIMIC-Location** and MIMIC-SubpopBench, and 500 steps on **SHAC**, **HateSpeech**, and Civilcomments. We checkpoint 10 times for model selection across all experiment settings. We double the steps for two-stage training algorithms, including JTT, DFR, and DualFilter. For the MultiNLI dataset, we concatenate the embedding of premise, hypothesis, their difference, and their product, aligning with Williams et al. (2018). All training jobs were distributed across two nodes, with the total experimentation consuming approximately 1,100 GPU hours.

E.5. Evaluation Metrics

DeconDTN-Toolkit supports the calculation of accuracy, f1 score, and Area Under the Precision-Recall Curve (AURPC), and Expected Calibration Error (ECE). These metrics are calculated in provenance-specific, micro-averaged, macro-averaged, and worst levels.

Appendix F. Infrastructure

F.1. Ablation study on provenance-balanced minibatches

We intended to follow DomainBed and Subpop-Bench to report the ERM with balanced minibatches, which stabilizes the optimization without altering the learning objective. An ablation study was conduct on the provenance-balanced-minibatch technique to investigate its effects on downstream performance. The ERM without balanced minibatches results in OOD WGA of 0.58, 0.42, 0.30, 0.33, 0.85 using default hyperparameters on the Civilcomments, **HateSpeech**, MIMIC-SubpopBench, MultiNLI, and **SHAC** datasets, respectively, falling within ranges of the reported ERM.

F.2. Training Speed

Using experiment settings of NVIDIA A100, BERT (`bert-base-uncased`), a batch size of 32 per provenance, and a maximum sequence length of 256, we derived the averaged step time across available datasets throughout the training process (Table 5).

Table 5: Mean and standard deviation of step time per algorithm.

Algorithm	Step Time (s)
CAD	0.62 (0.19)
CDANN	0.59 (0.20)
CORAL	0.60 (0.21)
DANN	0.59 (0.21)
DFR	0.22 (0.08)
DownSampling	0.57 (0.22)
DualFilter	0.62 (0.17)
ERM	0.58 (0.21)
Fish	1.34 (0.46)
GroupDRO	0.61 (0.21)
IRM	0.61 (0.21)
JTT	0.78 (0.26)
LISA	0.59 (0.22)
LfF	1.27 (0.44)
MMD	0.62 (0.21)
MTL	0.62 (0.22)
Mixup	0.62 (0.22)
UpSampling	0.58 (0.22)

Appendix G. Full Results

G.1. Optimal Hyperparameter

We list the optimal hyperparameters per algorithm for random search in Table 6 and Table 7 for **Source Datasets** and Table 8 and Table 9 for Attribute Datasets.

G.2. Figures

In-distribution (solid) and OOD (dashed) WGA for ERM on Attribute Datasets are demonstrated in Figure 3. The relationship between OOD WGA vs. ID WGA and OOD WGA and *alpha* are on Attribute Datasets illustrated in Figure 4.

G.3. Full Metrics

We list OOD WGA across algorithms in Attribute Datasets in Table 10. Similar to WGA, we consider provenance in metric calculation. We list the micro-averaged and provenance-macro-averaged AUPRC per algorithm calculated at $\alpha^{te} = 0.6$ for **Source Datasets** in Table 11 and Table 12. Similarly, the micro-averaged and provenance-macro-averaged AUPRC for Attribute Datasets are listed in Table 13 and Table 14.

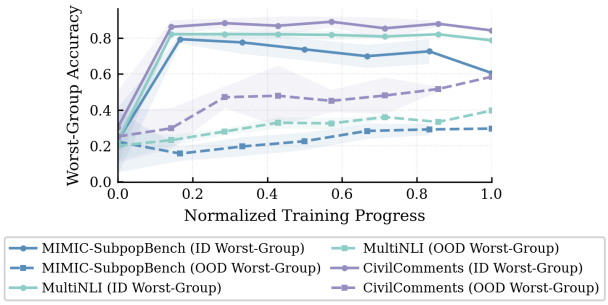

Figure 3: In-distribution (solid) and OOD (dashed) WGA for ERM in **Attribute Datasets**. Y-axis: Worst Group Accuracy (WGA). X-axis: normalized progress of training. The gap between the dashed and solid lines shows that models make inaccurate predictions in the OOD setting throughout their training process.

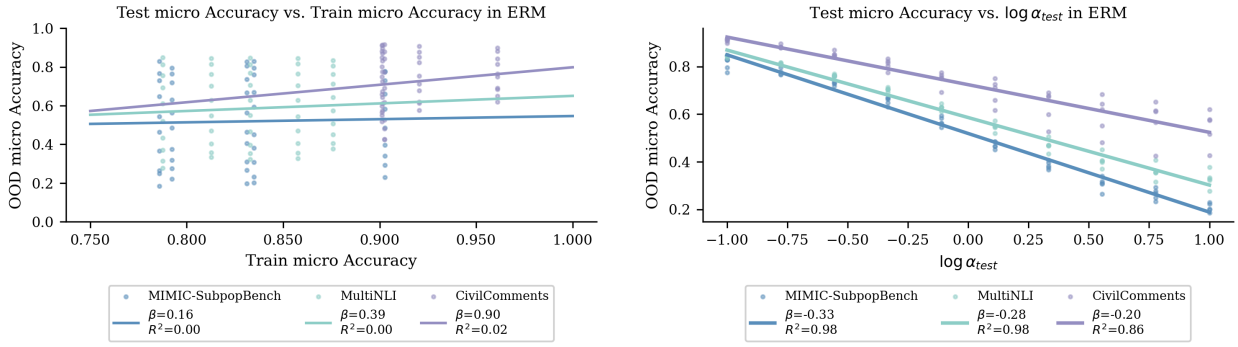

Figure 4: WGA is not "on the line" with respect to ID performance, but exhibits a strong linear relationship with the shift parameter α in **Attribute Datasets**.

Table 4: Hyperparameters, their default values and distributions for random search.

Condition	**Parameter**	**Default value**	**Random distribution**
all	learning rate	1e-5	$10^{\text{Uniform}(-5,-3.5)}$
	weight decay	0	$10^{\text{Uniform}(-6,-2)}$
DANN, CDANN	lambda	1.0	$10^{\text{Uniform}(-2,2)}$
	discriminator weight decay	0	$10^{\text{Uniform}(-6,-2)}$
	discriminator steps	1	$2^{\text{Uniform}(0,3)}$
	discriminator width	256	$2^{\text{Uniform}(6,10)}$
	discriminator depth	3	RandomChoice($[3, 4, 5]$)
	discriminator dropout	0	RandomChoice($[0, 0.1, 0.5]$)
	gradient penalty	0	$10^{\text{Uniform}(-2,1)}$
	adam β_1	0.5	RandomChoice($[0, 0.5]$)
IRM	lambda	100	$10^{\text{Uniform}(-1,5)}$
	iterations of penalty annealing	500	$10^{\text{Uniform}(0,4)}$
Mixup	alpha	0.2	$10^{\text{Uniform}(0,4)}$
LISA	alpha	2.0	$10^{\text{Uniform}(-1,1)}$
	intra-domain mixup ratio	0.5	Uniform$(0, 1)$
	mixup method	mixup	RandomChoice([mixup, cut mixup])
GroupDRO	eta	0.01	$10^{\text{Uniform}(-1,1)}$
MMD, CORAL	gamma	1	$10^{\text{Uniform}(-1,1)}$
Fish	meta learning rate	0.5	RandomChoice($[0.05, 0.1, 0.5]$)
MTL	exponential moving average	0.99	RandomChoice($[0.5, 0.9, 0.99, 1]$)
CAD	lambda	0.1	$10^{\text{RandomChoice}([-4,-2,-1,0,1,2])}$
	temperature	0.1	RandomChoice($[0.05, 0.1]$)
JTT	first stage fraction	0.5	$10^{\text{Uniform}(0.2,0.8)}$
	lambda l_2 penalty	0.1	$10^{\text{Uniform}(0,2.5)}$
DFR	first stage fraction	0.5	$10^{\text{Uniform}(0.2,0.8)}$
	second stage l_2 penalty	0.1	$10^{\text{Uniform}(-2,0.5)}$
LfF	amplification degree	0.1	Uniform$(0.05, 0.3)$
DualFilter	mask type	A	RandomChoice($[D, I, A]$)
	mask threshold	0.5	Uniform$(0.5, 0.9)$
	ablation rate	0.5	Uniform$(0.5, 0.9)$
	warm up steps	50	RandomChoice$(10, 25, 50)$
	embedding mask	True	RandomChoice([False, True])
	classifier mask	False	RandomChoice([False, True])

Table 6: Hyperparameters and Optimal Values in **Source Datasets**

Algorithm	Parameter	Optimal Value		
		HateSpeech	**MIMIC-Location**	**SHAC**
CAD	lambda	1.0e-03	1.0e-04	1.0e-03
	learning rate	1.6e-05	2.1e-05	2.2e-05
	temperature	0.1	0.1	0.1
	weight decay	3.5e-05	1.1e-03	1.0e-05
CDANN	adam β_1	0.5	0.5	0
	discriminator depth	3	3	3
	discriminator dropout	0.5	0	0.1
	discriminator learning rate	1.1e-05	5.0e-05	1.8e-04
	discriminator steps	2	1	1
	discriminator weight decay	1.5e-05	0	7.2e-04
	discriminator width	220	256	829
	generator learning rate	4.6e-05	5.0e-05	7.1e-05
	generator weight decay	1.4e-06	0	2.2e-03
	gradient penalty	0.5	0	0.0
	lambda	0.8	1.0	5.8
	lr	1.6e-05	1.0e-05	2.2e-05
	weight decay	3.5e-05	0	1.0e-05
CORAL	gamma	0.3	1.6	0.3
	learning rate	7.7e-05	1.4e-05	6.7e-05
	weight decay	1.8e-04	7.9e-06	2.8e-06
DANN	adam β_1	0.5	0.5	0.5
	discriminator depth	3	3	3
	discriminator dropout	0	0	0
	discriminator learning rate	5.0e-05	5.0e-05	5.0e-05
	discriminator steps	1	1	1
	discriminator weight decay	0	0	0
	discriminator width	256	256	256
	generator learning rate	5.0e-05	5.0e-05	5.0e-05
	generator weight decay	0	0	0
	gradient penalty	0	0	0
	lambda	1.0	1.0	1.0
	lr	1.0e-05	1.0e-05	1.0e-05
	weight decay	0	0	0
DFR	first stage fraction	0.7	0.6	0.3
	learning rate	6.1e-05	7.7e-05	2.1e-05
	second stage l_2 penalty	0.3	1.6	0.0
	weight decay	1.4e-05	1.8e-04	1.1e-03
DownSampling	learning rate	5.2e-05	4.4e-05	5.2e-05
	weight decay	2.8e-05	7.3e-03	2.8e-05
DualFilter	ablation rate	0.8	0.5	0.7
	classifier mask	0	False	0
	embedding mask	0	True	0
	first stage fraction	0.5	0.5	0.5
	learning rate	6.1e-05	1.0e-05	5.2e-05
	mask threshold	0.9	0.5	0.6
	mask type	I	A	A
	warm up steps	50	50	50
	weight decay	1.4e-05	0	2.8e-05

Table 7: Hyperparameters and Optimal Values in **Source Datasets** (Continued)

Algorithm	Parameter	Optimal Value		
		HateSpeech	**MIMIC-Location**	**SHAC**
ERM	learning rate	5.2e-05	2.2e-05	8.9e-06
	weight decay	2.8e-05	1.0e-05	2.2e-05
Fish	learning rate	6.7e-05	6.7e-05	6.7e-05
	meta learning rate	0.5	0.5	0.5
	weight decay	2.8e-06	2.8e-06	2.8e-06
GroupDRO	eta	1.0e-02	0.0	0.0
	learning rate	1.0e-05	7.7e-05	8.5e-06
	weight decay	0	1.8e-04	4.5e-04
IRM	iterations of penalty annealing	247	247	3211
	lambda	1.9	1.9	2.7e+02
	learning rate	6.1e-05	6.1e-05	2.1e-05
	weight decay	1.4e-05	1.4e-05	1.1e-03
JTT	first stage fraction	0.5	0.6	0.6
	lambda l_2 penalty	10.0	9.6	34.9
	learning rate	1.8e-05	7.7e-05	6.7e-05
	weight decay	5.1e-03	1.8e-04	2.8e-06
LISA	alpha	6.2	1.3	3.2
	intra-domain mixup ratio	0.9	0.0	1.0
	learning rate	6.7e-05	8.5e-06	2.2e-05
	mixup method	mixup	cutmix	cutmix
	weight decay	2.8e-06	4.5e-04	1.0e-05
LfF	amplification degree	0.1	0.1	0.1
	learning rate	1.0e-05	1.0e-05	8.9e-06
	weight decay	0	0	2.2e-05
MMD	gamma	0.1	0.1	0.4
	learning rate	4.4e-05	4.4e-05	1.4e-05
	weight decay	7.3e-03	7.3e-03	7.9e-06
MTL	exponential moving average	1.0	1.0	1.0
	learning rate	7.7e-05	2.2e-05	7.7e-05
	weight decay	1.8e-04	1.0e-05	1.8e-04
Mixup	alpha	0.6	0.3	3.1
	learning rate	2.2e-05	1.4e-05	6.1e-05
	weight decay	1.0e-05	7.9e-06	1.4e-05
UpSampling	learning rate	6.7e-05	6.1e-05	4.4e-05
	weight decay	2.8e-06	1.4e-05	7.3e-03

Table 8: Hyperparameters and Optimal Values in **Attribute Datasets**

Algorithm	Parameter	Optimal Value		
		Civilcomments	MIMIC-SubpopBench	MultiNLI
CAD	lambda	1.0e-03	1.0e-04	1.0e-03
	learning rate	1.6e-05	6.1e-05	4.2e-05
	temperature	0.1	0.1	0.1
	weight decay	3.5e-05	1.4e-05	1.7e-05
CDANN	adam β_1	0.5	0.5	0.5
	discriminator depth	3	5	4
	discriminator dropout	0	0.5	0.5
	discriminator learning rate	5.0e-05	1.2e-04	1.3e-05
	discriminator steps	1	1	2
	discriminator weight decay	0	5.2e-03	1.0e-04
	discriminator width	256	127	106
	generator learning rate	5.0e-05	6.8e-05	2.2e-04
	generator weight decay	0	1.3e-06	1.8e-06
	gradient penalty	0	0.1	0.2
	lambda	1.0	0.0	0.0
	lr	1.0e-05	9.1e-05	7.7e-05
	weight decay	0	6.2e-04	1.8e-04
CORAL	gamma	0.6	0.5	0.6
	learning rate	5.2e-05	2.1e-05	4.2e-05
	weight decay	2.8e-05	1.1e-03	1.7e-05
DANN	adam β_1	0.5	0.5	0.5
	discriminator depth	3	5	4
	discriminator dropout	0.5	0.1	0.5
	discriminator learning rate	1.1e-05	1.7e-05	1.3e-05
	discriminator steps	2	4	2
	discriminator weight decay	1.5e-05	4.2e-03	1.0e-04
	discriminator width	220	497	106
	generator learning rate	4.6e-05	6.0e-05	2.2e-04
	generator weight decay	1.4e-06	1.9e-03	1.8e-06
	gradient penalty	0.5	0.0	0.2
	lambda	0.8	4.9	0.0
	lr	1.6e-05	1.8e-05	7.7e-05
	weight decay	3.5e-05	5.1e-03	1.8e-04
DFR	first stage fraction	0.7	0.3	0.6
	learning rate	6.1e-05	2.1e-05	6.7e-05
	second stage l_2 penalty	0.3	0.0	0.0
	weight decay	1.4e-05	1.1e-03	2.8e-06
DownSampling	learning rate	4.2e-05	8.9e-06	9.9e-06
	weight decay	1.7e-05	2.2e-05	6.3e-06
DualFilter	ablation rate	0.7	0.8	0.8
	classifier mask	0	0	0
	embedding mask	0	0	0
	first stage fraction	0.5	0.5	0.5
	learning rate	5.2e-05	6.1e-05	6.1e-05
	mask threshold	0.6	0.9	0.9
	mask type	A	I	I
	warm up steps	50	50	50
	weight decay	2.8e-05	1.4e-05	1.4e-05

Table 9: Hyperparameters and Optimal Values in **Attribute Datasets** (Continued)

Algorithm	Parameter	Optimal Value		
		Civilcomments	MIMIC-SubpopBench	MultiNLI
ERM	learning rate	4.4e-05	8.9e-06	4.2e-05
	weight decay	7.3e-03	2.2e-05	1.7e-05
Fish	learning rate	9.1e-05	8.9e-06	1.6e-05
	meta learning rate	0.1	0.1	0.5
	weight decay	6.2e-04	2.2e-05	3.5e-05
GroupDRO	eta	0.0	0.0	0.0
	learning rate	9.9e-06	1.8e-05	6.7e-05
	weight decay	6.3e-06	5.1e-03	2.8e-06
IRM	iterations of penalty annealing	3001	3211	3775
	lambda	29.4	2.7e+02	7.5e+04
	learning rate	2.2e-05	2.1e-05	7.7e-05
	weight decay	1.0e-05	1.1e-03	1.8e-04
JTT	first stage fraction	0.3	0.7	0.8
	lambda l_2 penalty	2.9	1.3e+02	34.7
	learning rate	2.1e-05	6.1e-05	9.1e-05
	weight decay	1.1e-03	1.4e-05	6.2e-04
LISA	alpha	0.2	0.1	0.2
	intra-domain mixup ratio	0.1	0.9	0.6
	learning rate	4.2e-05	8.9e-06	5.2e-05
	mixup method	mixup	mixup	mixup
	weight decay	1.7e-05	2.2e-05	2.8e-05
LfF	amplification degree	0.1	0.1	0.1
	learning rate	8.5e-06	8.5e-06	1.4e-05
	weight decay	4.5e-04	4.5e-04	7.9e-06
MMD	gamma	0.1	0.1	0.1
	learning rate	4.4e-05	4.4e-05	4.4e-05
	weight decay	7.3e-03	7.3e-03	7.3e-03
MTL	exponential moving average	0.9	0.9	1.0
	learning rate	1.4e-05	6.7e-05	5.2e-05
	weight decay	7.9e-06	2.8e-06	2.8e-05
Mixup	alpha	0.1	0.6	0.2
	learning rate	8.9e-06	2.2e-05	7.7e-05
	weight decay	2.2e-05	1.0e-05	1.8e-04
UpSampling	learning rate	5.2e-05	1.0e-05	6.7e-05
	weight decay	2.8e-05	0	2.8e-06

Table 10: OOD WGA across algorithms in **Attribute Datasets**

Algorithm	**MIMIC-SubpopBench**	**MultiNLI**	**Civilcomments**	**Avg**
ERM	0.28 ± 0.02	0.36 ± 0.05	0.51 ± 0.05	0.38
UpSampling	0.46 ± 0.03	0.56 ± 0.02	0.67 ± 0.03	0.56
DownSampling	0.51 ± 0.05	0.63 ± 0.00	0.72 ± 0.02	0.62
CAD	0.30 ± 0.06	0.34 ± 0.02	0.58 ± 0.06	0.41
CDANN	0.31 ± 0.02	0.33 ± 0.04	0.50 ± 0.07	0.38
CORAL	0.30 ± 0.03	0.32 ± 0.04	0.53 ± 0.09	0.39
DANN	0.27 ± 0.02	0.31 ± 0.03	0.45 ± 0.04	0.34
DFR	0.37 ± 0.05	0.60 ± 0.01	0.58 ± 0.07	0.52
DualFilter	0.31 ± 0.06	0.40 ± 0.02	0.53 ± 0.05	0.41
Fish	0.30 ± 0.03	0.31 ± 0.01	0.49 ± 0.05	0.37
GroupDRO	0.28 ± 0.03	0.35 ± 0.04	0.52 ± 0.05	0.38
IRM	0.31 ± 0.04	0.40 ± 0.04	0.54 ± 0.07	0.42
JTT	0.28 ± 0.06	0.34 ± 0.07	0.52 ± 0.06	0.38
LISA	0.47 ± 0.02	0.64 ± 0.01	0.67 ± 0.05	0.60
LfF	0.32 ± 0.06	0.35 ± 0.04	0.53 ± 0.03	0.40
MMD	0.34 ± 0.03	0.35 ± 0.03	0.52 ± 0.05	0.40
MTL	0.34 ± 0.06	0.36 ± 0.04	0.53 ± 0.04	0.41
Mixup	0.29 ± 0.02	0.36 ± 0.05	0.53 ± 0.06	0.39

Table 11: OOD micro-averaged Area Under the Precision-Recall Curve in **Source Datasets**

Algorithm	**SHAC**	**MIMIC-Location**	**HateSpeech**	**Avg**
ERM	0.91 ± 0.02	0.48 ± 0.04	0.69 ± 0.06	0.69
UpSampling	0.96 ± 0.01	0.53 ± 0.03	0.76 ± 0.03	0.75
DownSampling	0.95 ± 0.01	0.54 ± 0.02	0.80 ± 0.02	0.76
CAD	0.92 ± 0.02	0.49 ± 0.03	0.66 ± 0.03	0.69
CDANN	0.95 ± 0.03	0.49 ± 0.02	0.66 ± 0.03	0.70
CORAL	0.95 ± 0.01	0.50 ± 0.03	0.70 ± 0.04	0.72
DANN	0.95 ± 0.02	0.50 ± 0.04	0.71 ± 0.03	0.72
DFR	0.93 ± 0.01	0.49 ± 0.02	0.69 ± 0.03	0.71
DualFilter	0.94 ± 0.02	0.49 ± 0.02	0.72 ± 0.03	0.72
Fish	0.93 ± 0.02	0.49 ± 0.03	0.66 ± 0.03	0.69
GroupDRO	0.90 ± 0.03	0.50 ± 0.02	0.67 ± 0.02	0.69
IRM	0.93 ± 0.02	0.50 ± 0.03	0.69 ± 0.05	0.70
JTT	0.92 ± 0.05	0.49 ± 0.03	0.67 ± 0.03	0.70
LISA	0.94 ± 0.04	0.52 ± 0.02	0.77 ± 0.01	0.74
LfF	0.87 ± 0.05	0.49 ± 0.02	0.61 ± 0.03	0.66
MMD	0.93 ± 0.02	0.47 ± 0.02	0.69 ± 0.02	0.70
MTL	0.92 ± 0.02	0.48 ± 0.02	0.64 ± 0.04	0.68
Mixup	0.91 ± 0.04	0.50 ± 0.02	0.71 ± 0.02	0.71

Table 12: OOD macro-averaged Area Under the Precision-Recall Curve on **Source Datasets**

Algorithm	**SHAC**	**MIMIC-Location**	**HateSpeech**	**Avg**
ERM	0.92 ± 0.03	0.57 ± 0.02	0.73 ± 0.03	0.74
UpSampling	0.96 ± 0.01	0.57 ± 0.03	0.75 ± 0.03	0.76
DownSampling	0.93 ± 0.02	0.54 ± 0.02	0.71 ± 0.01	0.73
CAD	0.94 ± 0.02	0.57 ± 0.02	0.74 ± 0.02	0.75
CDANN	0.94 ± 0.02	0.55 ± 0.02	0.72 ± 0.02	0.74
CORAL	0.91 ± 0.02	0.57 ± 0.02	0.72 ± 0.03	0.74
DANN	0.89 ± 0.08	0.54 ± 0.02	0.72 ± 0.03	0.72
DFR	0.94 ± 0.02	0.55 ± 0.03	0.75 ± 0.03	0.75
DualFilter	0.95 ± 0.01	0.57 ± 0.02	0.75 ± 0.04	0.76
Fish	0.94 ± 0.02	0.57 ± 0.02	0.72 ± 0.03	0.74
GroupDRO	0.93 ± 0.04	0.57 ± 0.02	0.73 ± 0.03	0.74
IRM	0.93 ± 0.01	0.57 ± 0.02	0.72 ± 0.01	0.74
JTT	0.93 ± 0.06	0.57 ± 0.03	0.73 ± 0.02	0.74
LISA	0.93 ± 0.05	0.55 ± 0.02	0.71 ± 0.02	0.73
LfF	0.91 ± 0.04	0.55 ± 0.01	0.69 ± 0.02	0.72
MMD	0.91 ± 0.03	0.55 ± 0.01	0.73 ± 0.01	0.73
MTL	0.90 ± 0.04	0.55 ± 0.03	0.67 ± 0.02	0.71
Mixup	0.84 ± 0.05	0.57 ± 0.03	0.72 ± 0.02	0.71

Table 13: OOD micro-averaged Area Under the Precision-Recall Curve on **Attribute Datasets**

Algorithm	**MIMIC-SubpopBench**	**MultiNLI**	**Civilcomments**	**Avg**
ERM	0.39 ± 0.02	0.49 ± 0.02	0.66 ± 0.10	0.51
UpSampling	0.48 ± 0.03	0.64 ± 0.02	0.83 ± 0.01	0.65
DownSampling	0.56 ± 0.03	0.72 ± 0.01	0.85 ± 0.02	0.71
CAD	0.39 ± 0.02	0.46 ± 0.02	0.71 ± 0.04	0.52
CDANN	0.40 ± 0.02	0.46 ± 0.02	0.68 ± 0.05	0.51
CORAL	0.40 ± 0.02	0.47 ± 0.03	0.71 ± 0.02	0.53
DANN	0.38 ± 0.01	0.48 ± 0.05	0.64 ± 0.05	0.50
DFR	0.45 ± 0.02	0.68 ± 0.01	0.71 ± 0.03	0.61
DualFilter	0.42 ± 0.03	0.53 ± 0.03	0.71 ± 0.02	0.55
Fish	0.38 ± 0.01	0.45 ± 0.02	0.60 ± 0.06	0.48
GroupDRO	0.37 ± 0.01	0.49 ± 0.03	0.65 ± 0.02	0.50
IRM	0.41 ± 0.02	0.50 ± 0.04	0.67 ± 0.03	0.53
JTT	0.42 ± 0.04	0.49 ± 0.04	0.66 ± 0.04	0.52
LISA	0.51 ± 0.04	0.72 ± 0.02	0.84 ± 0.03	0.69
LfF	0.40 ± 0.03	0.48 ± 0.03	0.64 ± 0.06	0.51
MMD	0.40 ± 0.01	0.48 ± 0.02	0.69 ± 0.03	0.52
MTL	0.42 ± 0.05	0.49 ± 0.04	0.63 ± 0.03	0.51
Mixup	0.39 ± 0.02	0.48 ± 0.03	0.67 ± 0.01	0.51

Table 14: OOD macro-averaged Area Under the Precision-Recall Curve on **Attribute Datasets**

Algorithm	**MIMIC-SubpopBench**	**MultiNLI**	**Civilcomments**	**Avg**
ERM	0.54 ± 0.03	0.67 ± 0.01	0.84 ± 0.05	0.69
UpSampling	0.56 ± 0.02	0.68 ± 0.01	0.84 ± 0.02	0.70
DownSampling	0.55 ± 0.01	0.69 ± 0.01	0.82 ± 0.03	0.69
CAD	0.53 ± 0.02	0.66 ± 0.01	0.84 ± 0.03	0.68
CDANN	0.55 ± 0.02	0.66 ± 0.01	0.83 ± 0.03	0.68
CORAL	0.54 ± 0.02	0.67 ± 0.01	0.84 ± 0.03	0.68
DANN	0.54 ± 0.01	0.67 ± 0.01	0.84 ± 0.02	0.68
DFR	0.57 ± 0.02	0.70 ± 0.01	0.84 ± 0.01	0.70
DualFilter	0.54 ± 0.03	0.69 ± 0.01	0.86 ± 0.01	0.69
Fish	0.53 ± 0.02	0.66 ± 0.01	0.79 ± 0.05	0.66
GroupDRO	0.52 ± 0.02	0.68 ± 0.01	0.83 ± 0.03	0.68
IRM	0.54 ± 0.02	0.67 ± 0.01	0.83 ± 0.02	0.68
JTT	0.54 ± 0.03	0.67 ± 0.01	0.83 ± 0.03	0.68
LISA	0.56 ± 0.01	0.71 ± 0.01	0.85 ± 0.03	0.71
LfF	0.53 ± 0.03	0.66 ± 0.01	0.79 ± 0.05	0.66
MMD	0.55 ± 0.01	0.67 ± 0.01	0.84 ± 0.02	0.69
MTL	0.53 ± 0.02	0.64 ± 0.03	0.75 ± 0.02	0.64
Mixup	0.53 ± 0.03	0.66 ± 0.02	0.83 ± 0.02	0.67