The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning

Anonymous authors
Paper under double-blind review

Abstract

Back in the early 20th century, a horse named Hans appeared to perform arithmetic and other intellectual tasks during exhibitions in Germany, while it actually relied solely on involuntary cues in the body language from the human trainer. Modern machine learning models are no different. These models are known to be sensitive to *spurious correlations* between non-essential features of the inputs (e.g., background, texture, and secondary objects) and the corresponding labels. Such features and their correlations with the labels are known as "spurious" because they tend to change with shifts in real-world data distributions, which can negatively impact the model's generalization and robustness. In this paper, we provide a comprehensive survey of this emerging issue, along with a fine-grained taxonomy of existing state-of-the-art methods for addressing spurious correlations in machine learning models. Additionally, we summarize existing datasets, benchmarks, and metrics to facilitate future research. The paper concludes with a discussion of the broader impacts, the recent advancements, and future challenges in the era of generative AI, aiming to provide valuable insights for researchers in the related domains of the machine learning community.

1 Introduction

While machine learning (ML) systems have achieved remarkable strides, distribution shifts and adversarial examples have become critical issues (Taori et al., 2020). It is increasingly evident that one of the major causes behind these issues is the severe reliance on spurious correlations between superficial features of the inputs (e.g., background, texture, and secondary objects) and the corresponding labels. When these correlations captured during training no longer hold in the test data, the performance of ML models deteriorates, leading to robustness issues and negative social impact in critical domains such as healthcare. For instance, studies in pneumonia detection with Convolutional Neural Networks (CNNs) have shown that the models often depend on extraneous cues, like metal tokens in chest X-ray scans from different hospitals, instead of learning the actual pathological features of the disease (Kirichenko et al., 2023; Zech et al., 2018). Therefore, addressing spurious correlations is of vital importance.

In recent years, spurious correlations have been referred to under various terms, such as shortcut learning, group robustness, and simplicity bias. Significant progress has been made in both detecting and mitigating these phenomena in fields such as computer vision (Wang et al., 2021; Yang et al., 2022b), natural language processing (Du et al., 2022b), and healthcare (Huang et al., 2022). Despite these advances, a comprehensive survey that formally defines spurious correlations and systematically reviews associated learning algorithms and challenges has yet to be published.

In this paper, we present the first comprehensive survey on spurious correlations, providing a formal definition, a taxonomy of current state-of-the-art methods for addressing spurious correlations in machine learning models, and an overview of relevant datasets, benchmarks, and evaluation metrics. We further discuss future research challenges and explore the potential role of foundation models in mitigating the adverse effects of spurious correlations. We hope this survey serves as a comprehensive reference and inspires further research in robust machine learning.

2 Spurious Correlations

In statistics, spurious correlation, namely "*correlations that do not imply causation*," refers to two variables that appear related to each other, but their true relationship is either coincidental or confounded by an external variable. This phenomenon leads to misleading or incorrect interpretations of data and models.

A classic historical analogy is *Clever Hans* (Figure 1), a horse in early 20th-century Germany that drew worldwide attention for performing arithmetic and other intellectual tasks (Lapuschkin et al., 2019). However, Hans did not actually understand these concepts; instead, he responded based on subtle, unintentional cues in his trainer's posture. Today, machine learning agents tend to exhibit similar behavior, relying on non-semantic, spurious features that correlate with the correct answer during training, but fail to generalize to broader or shifted distributions. We first give a formal definition.

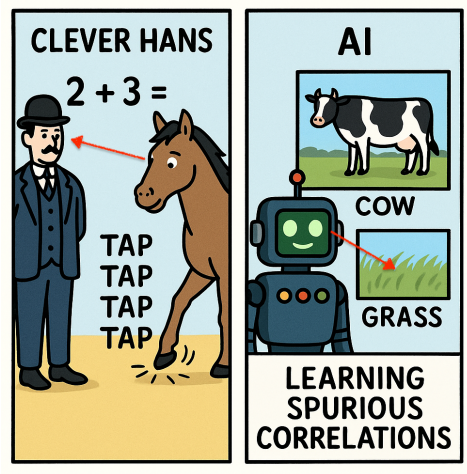


Figure 1: An illustration of the Clever Hans effect as an analogy for spurious correlations in machine learning. Just as Clever Hans appeared to solve arithmetic problems by responding to subtle cues from his trainer, machine learning models can achieve high accuracy by exploiting spurious features, e.g., associating grass with cows rather than learning true underlying concepts. (Image generated by GPT-4o)

Definition 2.1 (Spurious Correlation)**.** Let $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^{n}$ be the training set with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, where $\mathcal{X}$ denotes the set of all possible inputs and $\mathcal{Y}$ denotes the set of K classes. For each data point x_i with class label y_i, there exists a spurious attribute $a_i \in \mathcal{A}$ that is non-predictive of y_i, where $\mathcal{A}$ denotes the set of all possible spurious attributes. A spurious correlation, denoted as $\langle y, a \rangle$, is the association between $y \in \mathcal{Y}$ and $a \in \mathcal{A}$, where y and a exist in a one-to-many mapping $\phi : \mathcal{A} \mapsto \mathcal{Y}^{K'}$ conditioned on $\mathcal{D}_{tr}$ (with $1 < K' \leq K$). A set of data exhibiting the spurious correlation $\langle y, a \rangle$ is annotated with the group label $g = (y, a)$, where $\mathcal{G} := \mathcal{Y} \times \mathcal{A}$ is the set of combinations of class labels and spurious attributes.

For example, in the training set $\mathcal{D}_{tr}$, one might observe two spurious correlations, $\langle y, a \rangle$ and $\langle y', a \rangle$, meaning that the same spurious attribute a appears in two classes. A model trained on $\mathcal{D}_{tr}$ might simply use a to predict y, which would lead to incorrect predictions on samples with the correlation $\langle y', a \rangle$. This failure to generalize across environments due to reliance on a is visualized in Figure 2, where a spurious correlation between background and label (e.g., cows appearing on green grass) breaks under an environment shift (e.g., cows in the desert), while the true label remains unchanged.

2.1 Where do Spurious Correlations Come From?

Spurious correlations often stem from selection biases in datasets. Datasets with limited data are often underspecified, such that multiple plausible hypotheses can describe the data equally well; following Occam's Razor, the simplest hypothesis may be preferred even if it relies on spurious patterns. This tendency reflects a more general phenomenon known as simplicity bias, where models develop an over-reliance on highly available but less predictive features, as opposed to less available but highly predictive features (Hermann et al., 2023). Moreover, imbalanced group labels can lead to over-representation of certain groups, causing the model to depend on correlations that do not hold for minority groups. Finally, sampling noise—i.e., random variations inherent in collected samples—can cause the dataset to misrepresent the true data distribution, leading the model to interpret noise as a meaningful correlation.

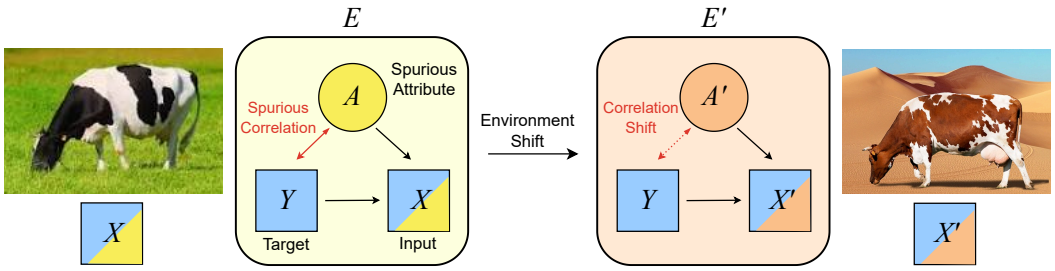


Figure 2: Depiction of a spurious correlation between spurious attribute A (e.g., grass) and target Y (e.g., cow). In environment E, input X contains both core features (invariant across environments) from Y, and spurious features from A. When shifting to a new environment E', the spurious attribute changes (e.g., desert background A'), causing the spurious correlation to break. The core features remain predictive of Y, but models trained on E may fail to generalize if they overly rely on A.

2.2 Why Are Machine Learning Models Sensitive to Spurious Correlations?

Inductive Biases from Learning Algorithms. The "no free lunch" theorem underscores that no single model is universally optimal and that each model incorporates an inductive bias (i.e., a set of assumptions about unseen data) that influences its predictions (Wolpert & Macready, 1997). For example, CNNs are inherently biased toward local connectivity and spatial invariance—properties that are suitable for image data. However, when the training data is biased, such inductive biases may lead the model to latch onto spurious patterns, resulting in overfitting and poor generalization.

Optimization. In the classical Empirical Risk Minimization (ERM) framework (Vapnik, 1991), let a sample (x, y, a) be drawn from a distribution conditioned on a group label $g = (y, a)$. Given a model with parameters θ, denoted $f_\theta : \mathcal{X} \to \mathbb{R}^K$, and a loss function $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ (e.g., softmax cross-entropy), the training objective is to minimize

$$\mathcal{L}_{\text{avg}}(f_\theta) := \mathbb{E}_{g \sim P_g} \mathbb{E}_{(x,y,a) \sim P_{x|g}} [\ell(f_\theta(x), y)]. \tag{1}$$

Because the optimization process does not take the spurious attribute a into account, if there is a strong correlation between y and a, the model may learn to rely on a when predicting y. In testing scenarios where the correlation does not hold, this reliance results in poor performance. It has been shown that models trained via ERM exhibit high worst-group errors, as measured by

$$\mathcal{L}_{\text{wg}}(f_\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y,a) \sim P_{x|g}} [\ell(f_\theta(x), y)]. \tag{2}$$

Methods such as Group DRO (Sagawa et al., 2019a) have been proposed to mitigate this effect by modifying the training objective to minimize the training loss within the worst-performing groups.

2.3 Theoretical Insights

Recent theoretical works have expanded our understanding of how spurious correlations emerge and persist within the latent space. Analyzing the latent space of a model can reveal how core and spurious features interact during training. The presence of spurious correlations alters not only what features are learned, but also the temporal dynamics of learning. Models often embed spurious features early during the training process due to their simplicity or strong correlation with the label. If we denote the learned representation of an input x as $\phi(x)$, then during early training, the model's prediction for class y often satisfies:

$$\langle \phi(x), w_y \rangle \approx \langle \phi(a), w_y \rangle \gg \langle \phi(x_c), w_y \rangle, \tag{3}$$

assuming our input x can be decomposed into core features x_c and spurious feature a, i.e., $x = (x_c, a)$. w_y is the classifier weight vector for class y. The model's prediction relies on a instead of x_c when minimizing

$\mathcal{L}_{\text{avg}}$ because a provides a more immediately predictive signal for y in the training distribution. Not only does this delay the learning of core features, remnants of $\phi(a)$ are retained in the representation space even as generalizable features are eventually learned (Qiu et al., 2024). Late-stage learning can even reinforce simplified representations that encode spurious patterns (Tsoy & Konstantinov, 2024).

Building on this, Bombari & Mondelli (2023) demonstrates that the degree of memorization of a spurious correlation can be geometrically characterized by the feature alignment between the spurious pattern and the training sample within the learned representation. Subsequent studies have challenged the reliability of post-hoc mitigation strategies. For instance, in two-agent rationale-predictor settings, spurious cues can originate from the rationale generator rather than the dataset (Liu et al., 2025a), while the effectiveness of regularization methods highly depends on the alignment between spurious attributes and both known and unknown concepts, as regularizers can suppress useful signals when concept correlations are entangled (Hong et al., 2025). Together, these works characterize spurious correlations as a structural and persistent phenomenon within the learned representation that indirectly results from the training objective.

3 Related Areas

We provide an overview of research fields closely related to spurious correlations, offering a broader context and complementary perspectives. **Domain Generalization** is a wider goal that aims to improve model performance on unseen distributions, and it often fails when models rely on spurious correlations that do not hold across domains. **Group Robustness** focuses on ensuring consistent performance across subgroups, and spurious correlations that vary between groups can cause severe performance drops in minority groups. **Shortcut Learning** arises when models exploit spurious correlations as easy-to-learn signals instead of capturing the underlying causal features. **Simplicity Bias** leads models to favor spurious correlations because they often correspond to low-complexity patterns that are easier to optimize. We discuss each term as follows.

Domain Generalization. Also known as out-of-distribution (OOD) generalization, domain generalization aims to train models on one or more related domains such that they generalize well to unseen test domains (Wang et al., 2023). However, spurious correlations are a significant threat to domain generalization (Yang et al., 2022a). The core challenge in DG arises because models can easily learn features that are predictive in the source domain but become unreliable or irrelevant in a new, unseen target domain. These learned, non-causal associations are spurious correlations.

The relationship between DG and spurious correlation is *inverse* and *adversarial*. The success of any DG method is contingent on its ability to avoid or mitigate spurious correlations. For instance, in Multi-Source Domain Generalization (MSDG) (Li et al., 2018; Ahuja et al., 2021; Gulrajani & Lopez-Paz, 2020; Zhu et al., 2024), a model might learn a feature that happens to be correlated with a label across several, but not all, possible domains. While this feature may improve performance on the available source data, it is a spurious shortcut that will likely fail on a target domain where the correlation does not exist. The problem is even more acute in Single Domain Generalization (SDG) (Li et al., 2021; Wan et al., 2022; Ilke et al., 2022; Guo et al.), where the model has even less data diversity to learn from, making it highly susceptible to latching onto spurious features specific to that one domain.

Group Robustness. Group robustness methods seek to ensure a model performs consistently well across different predefined groups or subpopulations (Yang et al., 2023c) in the data, rather than achieving high overall accuracy at the expense of minority groups (Sagawa et al., 2019a). This concept is closely related to fairness and worst-case optimization: if a model latches onto a spurious feature that is prevalent in a majority group, its performance can degrade sharply on a minority group where that correlation does not hold. Spurious correlations often manifest as such group-dependent performance gaps. For instance, in the Waterbirds dataset (Sagawa et al., 2020b), a bird classification model may learn to associate "water background" with the label waterbird (since most training images of waterbirds have water in the background), which leads to poor accuracy on the minority group of waterbirds on land. Group robustness techniques address this by actively discouraging the model from exploiting these group-specific shortcuts. One prominent approach is Group Distributionally Robust Optimization (Group DRO), which minimizes the worst-case loss over all groups rather than the average loss (Sagawa et al., 2020b). By focusing on the worst-performing group

during training, the model is pushed to learn features that work for all groups, thereby reducing its reliance on spurious cues present only in the majority group. Follow-up research has proposed various improvements, such as up-weighting under-represented group examples or using two-stage training to identify and then rectify bias against minority groups (Liu et al., 2021; Idrissi et al., 2022). Notably, benchmarks like WILDS (Koh et al., 2020) include several real-world spurious correlation tasks (e.g., Waterbirds, CelebA hair color classification, and CivilComments toxicity classification) explicitly designed to evaluate group robustness – measuring success by the model's accuracy on the worst-group (e.g., waterbirds on land, images of a certain attribute in the minority context). By treating spurious correlation issues as a group shift problem, group robustness frameworks provide effective tools to diagnose and improve a model's worst-case behavior under distribution shifts.

Shortcut Learning. Shortcut learning refers to this phenomenon where models rely on spurious correlations (often easier to learn) as opposed to the more relevant but complex features. These shortcuts are essentially spurious correlations that hold in the training set. For example, a vision model might classify images by background or texture instead of recognizing the animal itself, a classic Clever Hans behavior that breaks when the context changes. Empirical studies have shown that ImageNet-trained convolutional networks tend to prefer texture over object shape as a recognition cue, which is one type of shortcut. This bias leads to poor generalization on shape-distorted images or domain-shifted images (Geirhos et al., 2020). In NLP, a model might learn to rely on the presence of specific keywords as a shortcut for sentiment or entailment, rather than truly understanding language, causing errors on examples where those keywords appear in an unrelated context (Du et al., 2022a). Shortcut learning is particularly pernicious under standard ERM training because the learner will gravitate to any discriminative feature that improves training accuracy, without regard for causal relevance. Research in this area often evaluates models on "counterfactually" adjusted or stress-test datasets. For instance, testing the camel vs. cow classifier on images where camels appear on grass or cows on sand reveals the shortcut reliance. Approaches to mitigate shortcut learning overlap with those for spurious correlation in general, including data augmentation to break the spurious association, specialized training objectives, or interpretability-driven methods to detect when a model focuses on the wrong features. In summary, shortcut learning encapsulates the model's temptation to "cheat" by using easy correlations, underscoring why robust evaluation beyond the i.i.d. test set is crucial for detecting spurious reasoning.

Simplicity Bias. Simplicity bias (SB) is the tendency of deep neural networks (especially when trained with standard gradient-based optimization) to preferentially learn simple patterns or heuristics first, often to the exclusion of more complex but relevant features (Shah et al., 2020). In other words, given multiple predictive signals in the data, networks biased by simplicity will latch onto the easiest-to-fit signal (e.g., a low-level texture or an easily segregated background color) before considering more intricate structures (like shape combinations or abstract relations). This bias provides a fundamental reason why spurious correlations can dominate a model's predictions: if a spurious feature offers a simpler decision rule that fits the training data (even if it's not fundamentally causal), the model is likely to adopt it due to SB. Shah et al. (2020) demonstrates that neural nets can indeed entirely ignore more informative features and instead rely on a less informative but simpler feature, resulting in poor generalization whenever the simple feature's correlation with the label changes. In the context of spurious correlations, SB means the model might, for example, focus on background scenery to classify objects because backgrounds are simpler to learn (roughly uniform for each class), even if object-specific features would be more reliable outside the training distribution (Tiwari & Shenoy, 2023). This predisposition is closely linked with the model's vulnerability to distribution shifts and even adversarial perturbations: a network that heavily relies on a few simple features can be brittle, as small input changes can disrupt those features. Recent work has begun to address simplicity bias by altering training curricula or model architectures to force the learning of more complex features (Tiwari & Shenoy, 2023). Nonetheless, SB remains a double-edged sword. It may help find a decision rule consistent with training data (an Occam's Razor effect), but it often aligns tightly with spurious correlations, thereby undermining robustness. Recognizing and mitigating simplicity bias is thus important for developing models that resist spurious correlations, ensuring that they do not overly prefer "easy" but misleading patterns at the cost of true generalization.

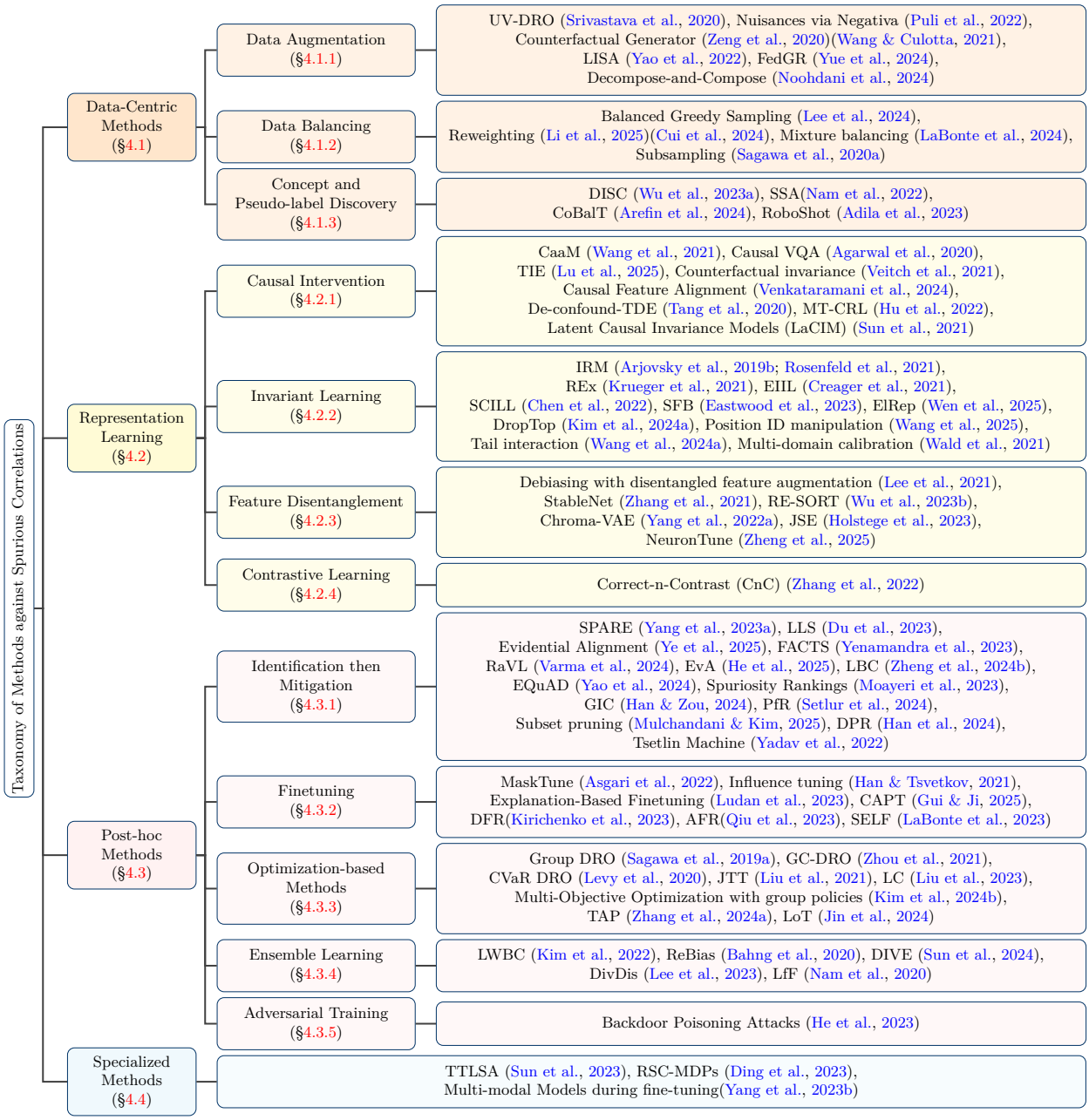


Figure 3: A comprehensive taxonomy of approaches to address spurious correlation in machine learning.

4 Methods

In this section, we discuss how various machine learning techniques address the issue of spurious correlation. To provide a clear framework, we organize these methods based on their stages in the machine learning model training pipeline: **Data-Centric Methods** (Section 4.1), which modify the training data itself; **Representation Learning** (Section 4.2) approaches, which alter the model's training objective or architecture to learn more robust features; **Post-hoc Methods** (Section 4.3), which are applied to correct or refine an already trained model; and **Specialized Methods** (Section 4.4) tailored to non-standard settings. Figure 3 illustrates a comprehensive taxonomy of the surveyed approaches.

4.1 Data-Centric Methods

Data-centric methods address spurious correlations at their source: the training data itself. Based on the principle that biases are learned from flawed or imbalanced data distributions, these techniques aim to modify the input data before or during training to break the harmful associations between spurious attributes and labels. This category encompasses a range of strategies, from **Data Augmentation** (Section 4.1.1) techniques that create new examples to expose the model to more diverse contexts, to **Data Balancing** (Section 4.1.2) methods that correct for distributional imbalances, such as reweighting and subsampling. Other approaches focus on **Concept and Pseudo-label Discovery** (Section 4.1.3) to explicitly identify and manage spurious attributes using labeled or unlabeled data.

4.1.1 Data Augmentation

Data augmentation increases the diversity of a dataset without the need for new data collection. Techniques such as image rotation, cropping, and noise injection are used to produce varied training samples, thereby improving model generalization by mitigating the impact of spurious correlations. For example, UV-DRO (Srivastava et al., 2020) leverages human annotations to augment training examples with potential unmeasured variables, reducing the spurious correlation problem to a covariate shift problem. When additional annotations are unavailable, methods such as Nuisances via Negativa (Puli et al., 2022) corrupt semantic information to highlight nuisances, while LISA (Yao et al., 2022) employs a mixup-based technique for selective augmentation. In graph settings, FedGR (Yue et al., 2024) introduces anti-shortcut augmentations by partitioning the graph into rationale and non-rationale subgraphs, perturbing the latter to dissociate spurious cues from the target signal, while also leveraging differences between local and global models within a federated learning paradigm.

Another class of augmentation approaches relies on generating counterfactual examples, or modified training samples that isolate causal and non-causal factors. This trains the model to ignore irrelevant details by separating spurious and true signals. The Counterfactual Generator (Zeng et al., 2020) creates counterfactual examples by intervening on existing observations, while Wang & Culotta (2021) proposes training a robust text classifier by automatically generating counterfactual training data. Decompose-and-Compose (Noohdani et al., 2024) takes an alternative approach, deconstructing images into causal and non-causal segments before recombining them into new samples.

4.1.2 Data Balancing

Spurious correlations are often caused by distributional imbalances in the training data, where spurious features consistently co-occur with target labels despite having no causal relationship. These issues can be addressed across various domains by modifying how existing data is sampled or weighted during training. For example, Balanced Greedy Sampling (Lee et al., 2024) is found to prevent such biases in continual learning tasks by retraining the final layer of the model using a balanced training subset. Reweighting has been used to mitigate dataset bias resulting from dataset condensation (Cui et al., 2024) and shortcut bias in multimodal settings (Li et al., 2025). However, LaBonte et al. (2024) warns that common group-balancing techniques can fail in specific scenarios: mini-batch upsampling and loss upweighting can observe a decrease in worst-group accuracy in later training epochs, and the effectiveness of data pruning varies depending on the group structure. Instead, they find that a combination of the two methods achieves a higher worst-group accuracy than each method individually. Similarly, Sagawa et al. (2020a) found that on overparameterized models, the subsampling on the majority group of data brings lower worst-group error instead of upweighting the minority group.

4.1.3 Concept and Pseudo-label Discovery

Another line of work in data manipulation involves predefining concepts and generating pseudo-labels to enhance the conceptual structure of the training data. This process helps to break intrinsic spurious correlations and improve model robustness. For instance, DISC (Wu et al., 2023a) discovers unstable predefined concepts across different environments, which are then used to augment the training data. Similarly, SSA (Nam et al., 2022) generates pseudo-labels for spurious attributes using a limited amount of annotated data,

and the pseudo-labeled attributes are then combined with the training data to train a robust model using worst-case loss minimization. Recent methods have explored unsupervised or zero-shot discovery of latent concepts to avoid the need for human labeling. CoBalT (Arefin et al., 2024) uses various clustering techniques to identify concepts, while RoboShot (Adila et al., 2023) uses a language model to generate concept subspaces by extracting the embedding of the task description from within a language model.

4.2 Representation Learning

In contrast to data-centric approaches, representation learning methods shift the focus from the data itself to the model's internal learning process and the quality of its learned features. By modifying the training objective or model architecture, these approaches allow the model to identify features that are inherently robust and causally linked to the outcome, rather than just statistically correlated. Key strategies include **Causal Intervention** (Section 4.2.1), which explicitly models the data's causal structure, and **Invariant Learning** (Section 4.2.2), which seeks features that remain stable across diverse environments. Other techniques focus on altering the latent space directly, such as **Feature Disentanglement** (Section 4.2.3), which aims to separate spurious and core features, and **Contrastive Learning** (Section 4.2.4), which enforces representational similarity for inputs that differ only in their spurious attributes.

4.2.1 Causal Intervention

Causal intervention methods focus on explicitly addressing the causal relationships between the input, label, and potential spurious attributes, aiming to improve model robustness and fairness by reducing the influence of spurious features. For example, CaaM (Wang et al., 2021) integrates a Causal Attention Module within the Vision Transformer architecture, using unsupervised learning to self-annotate and mitigate confounding effects. Similarly, Causal VQA (Agarwal et al., 2020) applies a semantic editing-based approach to assess and improve the robustness of Visual Question Answering (VQA) models, and TIE (Lu et al., 2025) leverages a translation operation within image embeddings based on text embeddings along a given spurious vector. Venkataramani et al. (2024) proposes Causal Feature Alignment, which uses a trained ERM classifier to extract core features within an image. Veitch et al. (2021) and Sun et al. (2021) incorporate structured causal graphs to reflect true relationships in the data, while De-confound-TDE (Tang et al., 2020) employs causal modeling to pinpoint the effect of momentum in long-tailed classification tasks. In multi-task settings, Multi-Task Causal Representation Learning (MT-CRL) (Hu et al., 2022) identifies the causal structure between different tasks to regularize the learning process, preventing the model from relying on spurious features that are shared across non-causally related tasks. In a similar vein, Veitch et al. (2021) introduces counterfactual invariance to evaluate the causal structure captured within the model by measuring whether predictions are influenced when changing irrelevant portions of a sample.

4.2.2 Invariant Learning

Invariant learning methods aim to train models to identify and focus on features that remain stable across different training environments, based upon the assumption that truly predictive features are invariant to environmental shifts. This strategy is supported by Wald et al. (2021), which establishes that models achieving multi-domain calibration are provably free of spurious correlations. Early work such as IRM (Arjovsky et al., 2019b) formulates an objective to learn deep invariant features that capture complex relationships between latent variables. Methods like REx (Krueger et al., 2021) and EIIL (Creager et al., 2021) further this goal by focusing on risk-level consistency and automatic environment inference, respectively, and SCILL (Chen et al., 2022) establishes theoretical group criteria to ensure group-invariant learning. Recent works emphasize enforcing invariance at the feature level. SFB (Eastwood et al., 2023) selectively amplifies features that are stable across domains. ElRep (Wen et al., 2025) introduces nuclear and Frobenius norm penalties on the representation matrix to balance sparsity and smoothness, improving generalization while avoiding over-regularization. DropTop (Kim et al., 2024a) dynamically identifies and debiases shortcut features in continual learning by selecting top-k activations that correlate with spurious behavior. In the LLM setting, Wang et al. (2025) demonstrates how manipulating position encodings can eliminate hidden shortcuts in role separation tasks, reinforcing invariant structure. Wang et al. (2024a) disentangles spurious and beneficial

correlations by combining frequency-restriction and interaction-based modules, enhancing the model's ability to learn domain-invariant representations.

4.2.3 Feature Disentanglement

Feature disentanglement methods aim to separate spurious representations from general representations in the latent space. These approaches often use dual-branch architectures. For example, StableNet (Zhang et al., 2021) removes both linear and non-linear dependencies using random Fourier features combined with a standard classifier, while Lee et al. (2021) trains two encoders to extract bias and intrinsic features separately. Disentanglement can be further enhanced through high- or low-dimensional projections. RE-SORT (Wu et al., 2023b) utilizes a Laplacian kernel function to project feature interactions to a higher dimension from within a multilevel stacked recurrent structure before targeted elimination through sample reweighting. In contrast, Chroma-VAE (Yang et al., 2022a) employs a dual-pronged VAE to disentangle latent representations in a low-dimensional subspace, ensuring that shortcut features do not dominate the learning process, while Holstege et al. (2023) jointly identifies two low-dimensional orthogonal subspaces within a vector representation to locate encoded spurious features prior to removal (JSE). Similarly, NeuronTune (Zheng et al., 2025) identifies the neurons in latent space that are affected by spurious correlations, and zeros out these neurons during the retraining process.

4.2.4 Contrastive Learning

Contrastive learning, a popular approach within self-supervised learning, has shown promise in addressing spurious correlations. The Correct-n-Contrast (CnC) method (Zhang et al., 2022) first trains an ERM model to identify samples within the same class that vary in spurious features, and then uses contrastive learning to encourage similar representations for these samples. The resulting model is better able to distinguish between essential and non-essential features.

4.3 Post-hoc Methods

Post-hoc methods are designed to mitigate spurious correlations in models that have already been trained, or during a secondary training stage like fine-tuning. Rather than building a robust model from scratch, these techniques typically operate on a pre-existing, potentially biased model, aiming to diagnose and correct its reliance on spurious features without full retraining. This diverse category includes **Identification then Mitigation** (Section 4.3.1) strategies that first find bias-conflicting samples and then reduce their influence; **Finetuning** (Section 4.3.2) approaches that refine a pre-trained model on carefully selected or transformed data; and **Optimization-based Methods** (Section 4.3.3) like Group DRO that modify the training objective to improve worst-group performance. Other prominent techniques involve **Ensemble Learning** (Section 4.3.4) to combine multiple biased models for a more robust prediction, and **Adversarial Training** (Section 4.3.5) to improve model resilience.

4.3.1 Identification then Mitigation

A number of methods follow an "identification then mitigation" strategy. These approaches first identify samples that are likely affected by spurious correlations and then mitigate their impact. SPARE (Yang et al., 2023a) and LLS (Du et al., 2023) both identify spurious correlations based on signs of simplicity bias, and use importance sampling or reweighting to reduce bias. Evidential Alignment (Ye et al., 2025) utilizes second-order risk minimization to identify and quantify the spurious correlations in pretrained models, then tunes the biased model based on overconfident samples from a calibration dataset. Spurious correlations can also be identified in the model's representations through latent feature probing. FACTS (Yenamandra et al., 2023) amplifies correlations to fit a bias-aligned hypothesis and then uses slicing via mixture modeling in the corresponding feature space to address underperforming data slices. RaVL (Varma et al., 2024) creates a region-aware loss function based on local image features, and EvA (He et al., 2025) learns class-specific spurious indicators within the model. LBC (Zheng et al., 2024b) and EQuAD (Yao et al., 2024) both quantify the likelihood and strength of spurious features through projection into a spurious embedding space and into a low-dimensional latent space, respectively. Mulchandani & Kim (2025) finds that spurious

correlations result from a small subset of the data and introduces a novel pruning technique that identifies and removes such samples. Disagreement Probability-based Resampling (DPR) (Han et al., 2024) takes the reverse approach, identifying and upsampling training examples that do not align with spurious correlations. In addition, Yadav et al. (2022) transforms natural language data into rule-based logic formulations and employs logical negation via a Tsetlin Machine to achieve explainable debiasing. Other methods like GIC, Spuriosity Rankings, and PfR infer group membership using auxiliary models or heuristic signals. GIC trains a classifier to predict group labels from spurious-label correlations, PfR uses zero-shot vision-language model predictions to identify spurious features, and Spuriosity Rankings proxies spuriosity using interpretable neural features to rank examples within each class and fine-tune on less biased data (Han & Zou, 2024; Moayeri et al., 2023; Setlur et al., 2024).

4.3.2 Finetuning

Finetuning strategies focus on refining a pre-trained general model by selectively adjusting sections of either the model or the dataset to reduce reliance on spurious features. For instance, MaskTune (Asgari et al., 2022) encourages the model to consider a wider array of input variables by mapping them to the same target, while Influence Tuning (Han & Tsvetkov, 2021) backpropagates attribution information for targeted refinement. In the text domain, Explanation-Based Finetuning (Ludan et al., 2023) trains models on artificially constructed datasets containing spurious cues, then tests on clean sets. Similarly, Causality-Aware Post-Training (CAPT) (Gui & Ji, 2025) fine-tunes models on a transformed dataset where specific events are first identified and then replaced with randomized, abstract symbols to break spurious correlations acquired during pre-training. Moreover, last-layer finetuning strategies such as Deep Feature Reweighting (DFR) (Kirichenko et al., 2023; Izmailov et al., 2022), Automatic Feature Reweighting (AFR) (Qiu et al., 2023), and Selective Last-Layer Finetuning (SELF) (LaBonte et al., 2023) have been shown to improve worst-group performance with minimal group annotations. Shuieh et al. (2025) compares various fine-tuning techniques, finding Supervised Fine-tuning to excel in complex, context-intensive tasks, while Direct Preference and Kahneman-Tversky Optimization both perform well in mathematical reasoning tasks.

4.3.3 Optimization-based Methods

Optimization-based strategies modify the training objective to promote better performance under spurious correlations. Group DRO (Sagawa et al., 2019a) minimizes the worst-case loss across different groups, ensuring robust performance across subpopulations. GC-DRO (Zhou et al., 2021) extends group DRO by jointly reweighting groups and individual instances (without needing perfectly clean group partitions or a noise model), thereby overcoming group DRO's failure under imperfect or mismatched groupings. CVaR DRO (Levy et al., 2020) extends this idea by optimizing the expected loss over the worst-case data distribution. Just Train Twice (JTT) (Liu et al., 2021) identifies and upsamples informative training samples with high training loss. Additionally, Logit Correction (LC) (Liu et al., 2023) adjusts model output logits based on group membership to balance performance across groups. In a similar vein, Kim et al. (2024b) introduces a Multi-Objective Optimization function that minimizes groupwise losses before adjusting weights to resolve conflicting objectives. Within deep CNNs, TAP (Zhang et al., 2024a) prevents spurious correlation reemergence by penalizing activations in later layers. LoT (Jin et al., 2024) provides a unique knowledge distillation training loop, where a teacher model trains student models to provide feedback to the main model, helping it capture invariant generalizations.

4.3.4 Ensemble Learning

Ensemble learning methods combine multiple independently trained biased models to produce an overall debiased prediction. For instance, LWBC (Kim et al., 2022) uses a committee of classifiers to identify bias-conflicting data and then emphasizes these samples during the main classifier's training. Similarly, ReBias (Bahng et al., 2020) trains de-biased representations by enforcing statistical independence from intentionally biased representations. DIVE (Sun et al., 2024) trains a collection of models on diverging subgraphs to encourage learning distinct patterns from the overall predictive graph. Some methods, such as DivDis (Lee et al., 2023) and LfF (Nam et al., 2020), adopt a two-stage ensemble strategy to identify and mitigate reliance on spurious features.

4.3.5 Adversarial Training

Adversarial training methods strengthen the model against adversarial inputs, thereby reducing reliance on spurious correlations. For instance, methods addressing backdoor poisoning attacks consider the spurious correlations introduced by watermarked, mislabeled training examples and propose mitigation strategies based on adversarial training (He et al., 2023).

4.4 Specialized Methods

This category includes approaches designed for specific settings. Test-Time Label-Shift Adaptation (TTLSA) (Sun et al., 2023) addresses label shifts at test time, and RSC-MDP (Ding et al., 2023) tackles spurious dependencies in reinforcement learning contexts. Additional work (Yang et al., 2023b) explores the use of multi-modal models to explicitly separate spurious attributes from the main class.

5 Datasets and Metrics

5.1 Datasets

This subsection provides an overview of benchmark datasets used for studying spurious correlations in machine learning, as summarized in Table 1. In these datasets, the correlation between the label y and a spurious attribute a observed during training is not guaranteed to hold during testing. The reviewed benchmarks, which are publicly available via the hyperlinks in Table 1, are categorized into five domains: Vision, Text, Graph, Multimodal, and Health. These datasets are constructed using both realistic and synthetic data, with synthetic generation being an increasingly prevalent approach (Qian et al., 2023). The distinct methodologies used to create both synthetic and realistic datasets are discussed below.

5.1.1 Synthetic Datasets

Synthetic datasets are created to enable a more controllable analysis of spurious correlations in machine learning models. Since spurious features do not appear systematically or easily isolated in standard benchmarks, manually synthesizing spurious features improves the efficiency of creating relevant datasets. Different types of spurious attributes could be manipulated to create challenging datasets targeting the model's robustness. In this subsection, common methodologies used to create synthetic datasets are categorized into four primary techniques: data composition, feature and noise injection, and style transfer.

Data Composition A primary method for synthetic dataset creation is the composition of data from multiple sources. For instance, to analyze the spurious correlation between a foreground object and its background, images from different datasets can be combined. An example of this is the UrbanCars dataset (Li et al., 2023), where the authors integrated foreground vehicle images from the StanfordCars dataset (Krause et al., 2013) with background scenes from the Places dataset (Zhou et al., 2017). Another common practice involves placing different object types onto a shared background. This technique is used to analyze potential spurious correlations between co-occurring objects or to study shortcut learning, as demonstrated in the ImageNet-W dataset (Li et al., 2023).

Pixel-Level Manipulation Manipulating existing datasets at the pixel level could also introduce desired spurious cues. While the target features still remain, the model might get biased towards the manipulation pattern. A notable example is the Colored MNIST dataset (Arjovsky et al., 2019a). In this work, the authors applied color filters to the original MNIST images to test whether a model learns the digit itself or relies on the color as a shortcut. In addition, noise injection is also a common method to introduce spurious attributes. In CIFAR-10-C and ImageNet-C, Hendrycks & Dietterich (2019) applied perturbations such as noise, blurring, weather filters, and image manipulations (e.g., brightness, contrast, etc.) to the CIFAR-10 and ImageNet datasets to evaluate the robustness of models. Similarly, in the SpuCoMNIST dataset (Joshi et al., 2023), varying levels of noise are injected into MNIST images to establish different spurious correlations.

Table 1: Datasets used to study spurious correlations in machine learning. Datasets are grouped by domain. The Synthetic column indicates whether a dataset is synthetic (✓) or derived from natural data (✗).

Dataset	Source Paper	#Class	#Attributes	#Sample	Brief Description	Synthetic	Domain
UrbanCars	Li et al. (2023)	2	2	8,000	Compositional vehicle scenes	✓	Vision
Colored MNIST	Arjovsky et al. (2019a)	–	2	60,000	Color-injected digits	✓	Vision
CIFAR-10-C	Hendrycks & Dietterich (2019)	–	15	60,000	Corruption on CIFAR-10	✓	Vision
SpuCo	Joshi et al. (2023)	4	4	118,100	Mixed synthetic + natural data	✓	Vision
Waterbirds	Sagawa et al. (2019b)	2	2	11,788	Composite birds + backgrounds	✓	Vision
ImageNet-C	Hendrycks & Dietterich (2019)	75	–	–	Corruption benchmark (Variant of ImageNet)	✓	Vision
ImageNet-R	Hendrycks et al. (2021a)	200	16	30,000	Artistic renditions (Variant of ImageNet)	✓	Vision
ImageNet-W	Li et al. (2023)	-	2	-	ImageNet with Watermark (Variant of ImageNet)	✓	Vision
ImageNet-9	Xiao et al. (2020)	9	–	5,495	Background-altered (Variant of ImageNet)	✓	Vision
Stylized-ImageNet	Geirhos et al. (2019)	–	–	–	AdaIN stylised variant (Variant of ImageNet)	✓	Vision
Hard ImageNet	Moayeri et al. (2022)	15	–	19,847	Spurious correlation specific (Subset of ImageNet)	✗	Vision
ImageNet-A	Hendrycks et al. (2021b)	200	–	7,500	Natural adversarial images (Subset of ImageNet)	✗	Vision
PACS	Li et al. (2017)	7	4	9,991	Multi-domain real images	✗	Vision
VLCS	Torralba & Efros (2011)	5	4	10,729	Aggregated real domains	✗	Vision
Office-Home	Venkateswara et al. (2017)	65	4	15,500	Multi-domain real images	✗	Vision
CelebA	Liu et al. (2015)	2	2	202,599	Face attributes	✗	Vision
MetaCoCo	Zhang et al. (2024b)	100	155	175,637	Few-shot data for spurious correlations	✗	Vision
BAR	Nam et al. (2020)	6	6	2,595	Biased action recognition	✗	Vision
NICO	He et al. (2021)	19	188	24,214	Context-shift animals	✗	Vision
MetaShift	Liang & Zou (2022)	410	–	12,868	Dataset-of-datasets shift	✗	Vision
FMOW	Christie et al. (2018)	63	–	1,047,691	Satellite imagery	✗	Vision
bFFHQ	Lee et al. (2021)	2	2	–	Balanced FFHQ faces	✗	Vision
CounterAnimal	Wang et al. (2024b)	45	14	13,100	Counterfactual animals	✗	Vision
WILDS-iWildCam	Koh et al. (2021)	182	–	203,029	Wildlife camera traps	✗	Vision
WILDS-GlobalWheat	Koh et al. (2021)	1	12	6,515	Wheat images	✗	Vision
WILDS-PovertyMap	Koh et al. (2021)	-	23×2	9,669	Satellite images	✗	Vision
QuAC	Choi et al. (2018)	–	–	98,407	Conversational QA	✗	Text
POVID	Zhou et al. (2024a)	–	–	17,000	VLM prompt–response	✗	Text
CivilComments	Borkan et al. (2019)	2	24	1,999,514	Online comments	✗	Text
MultiNLI	Williams et al. (2017)	3	2	206,175	Natural language inference	✗	Text
FDCL18	Founta et al. (2018)	7	–	80,000	Offensive speech	✗	Text
WILDS-Amazon	Koh et al. (2021)	-	2	539,502	Customer reviews on Amazon	✗	Text
ShortcutSuite	Yuan et al. (2024)	-	-	-	Shortcut in LLM	✗	Text
Spurious-Motif	Wu et al. (2022)	3	3	18, 000	Compositional motif-based graph	✓	Graph
GOOD-Motif	Gui et al. (2022)	5	3	30,000	Compositional motif-based graph	✓	Graph
GOOD-HIV	Gui et al. (2022)	2	2	41,127	Molecule dataset	✗	Graph
GOOD-SST2	Gui et al. (2022)	2	-	70,042	Sentence polarity	✗	Graph
DrugOOD	Ji et al. (2023)	–	5	58,000	Drug discovery	✗	Graph
IV/CV-VQA	Agarwal et al. (2020)	–	–	716,603	Causal VQA	✗	Multimodal
MM-SpuBench	Ye et al. (2024)	–	–	2,400	Multimodal spurious correlations	✗	Multimodal
ChestX-ray14	Wang et al. (2017)	14	–	112,120	Chest X-ray	✗	Health
MIMIC-CXR	Johnson et al. (2019)	14	–	377,110	Chest X-ray	✗	Health
CheXpert	Irvin et al. (2019)	14	–	224,316	Chest X-ray	✗	Health
PadChest	Bustos et al. (2020)	19	–	160,868	Chest X-ray	✗	Health
COVID-19	Cohen et al. (2020)	2	–	21,165	Chest X-ray	✗	Health

Style Transfer Machine learning models can exhibit biases toward object texture or style rather than shape. Style transfer is used to create datasets that test for this phenomenon. For example, the Stylized-ImageNet dataset (Geirhos et al., 2019) was generated by applying various artistic styles to images from ImageNet. The style transfer was performed using the AdaIN algorithm, which was trained on images from Kaggle's "Painter by Numbers" dataset (Huang & Belongie, 2017). The resulting dataset is instrumental in identifying whether a model exhibits a "shape bias" or a "texture bias".

5.1.2 Realistic Datasets

In contrast to synthetic datasets, which are engineered for controlled analysis, realistic datasets are curated from real-world sources to capture the natural, often subtle and unexpected, spurious correlations that models are likely to encounter in deployment. While they may not offer the same level of granular control as their synthetic counterparts, these benchmarks are indispensable for evaluating a model's performance on authentic data distributions. The curation of such datasets involves diverse methodologies, which we discuss below based on their primary data sourcing strategy.

Data Curation from Existing Sources A common strategy for creating new benchmarks is to curate and fuse data from existing sources, such as public datasets or search engines. This approach is prevalent in domain generalization, where datasets are constructed by combining data from different domains. For instance, the PACS dataset (Li et al., 2017) aggregates images of the same object classes from four distinct

domains, ranging from photorealistic sources like Caltech256 (Griffin et al., 2022) to sketch-based sources like TU-Berlin (Eitz et al., 2012). In other cases, datasets are curated by sub-sampling or re-annotating a single source. An example is bFFHQ (Lee et al., 2021), which was created by adding new annotations for a protected attribute to the original FFHQ dataset (Karras et al., 2019).

Collection from Scratch Beyond leveraging existing data, some benchmarks are constructed by collecting and annotating data directly from real-world environments. This approach is critical for capturing authentic data distributions and task-specific nuances that may be absent in pre-existing datasets. This methodology is applied across various fields: the FMOW dataset (Christie et al., 2018) contains satellite imagery for land use classification; the CivilComments dataset (Borkan et al., 2019) comprises user-generated comments for toxicity detection; and the PadChest dataset (Bustos et al., 2020) includes chest x-ray images and corresponding radiological reports from a hospital. A defining characteristic of these datasets is the significant manual annotation effort they require, which is essential for establishing ground-truth labels.

Mining Challenging Example A specialized curation method involves mining a large benchmark for examples that are inherently challenging for standard machine learning models. Unlike methods that select data based on class or domain, this technique specifically retrieves instances that are likely to elicit incorrect predictions due to strong spurious cues or other adversarial properties. For example, the Hard ImageNet dataset (Moayeri et al., 2022) consists of images from ImageNet (Russakovsky et al., 2015) that were identified as containing potent spurious features. Similarly, the ImageNet-A dataset (Hendrycks et al., 2021b) is composed of naturally occurring adversarial examples from ImageNet that consistently fool well-established classification models.

5.2 Metrics

A commonly used metric for quantifying robustness to spurious correlations is the **worst-group accuracy** $\mathrm{Acc}_{\mathrm{wg}}(f_\theta)$ (Idrissi et al., 2022; Chaudhuri et al., 2023), defined as the minimum accuracy across group-labeled test sets:

$$\mathrm{Acc}_{\mathrm{wg}}(f_\theta) \coloneqq \min_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\mathcal{D}^g_{\mathrm{test}}}\big[\mathbb{1}\big(f_\theta(x) = y\big)\big], \tag{4}$$

where $g = (y, a)$ is a group label determined by the target y and the spurious attribute a.

Other relevant metrics include:

Average-group accuracy. This measures the mean accuracy across all groups:

$$\mathrm{Acc}_{\mathrm{ag}}(f_\theta) \coloneqq \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\mathcal{D}^g_{\mathrm{test}}}\big[\mathbb{1}\big(f_\theta(x) = y\big)\big]. \tag{5}$$

Unlike worst-group accuracy, this metric is less sensitive to the hardest subgroup and instead reflects whether performance is balanced across groups on average.

Bias-conflicting accuracy. This is defined on a subset of groups $\mathcal{G}_{bc} \subseteq \mathcal{G}$ where the spurious attribute a does not align with the majority correlation observed during training:

$$\mathrm{Acc}_{\mathrm{bc}}(f_\theta) \coloneqq \frac{1}{|\mathcal{G}_{bc}|} \sum_{g \in \mathcal{G}_{bc}} \mathbb{E}_{(x,y)\sim\mathcal{D}^g_{\mathrm{test}}}\big[\mathbb{1}\big(f_\theta(x) = y\big)\big]. \tag{6}$$

This metric isolates model performance on the challenging bias-conflicting cases, where reliance on spurious features is most detrimental.

Finally, regular average accuracy across the entire test set remains useful for capturing overall utility, though it does not directly reflect robustness to spurious correlations. Together, these complementary metrics provide a more holistic view of a model's dependence on spurious features.

6 Broader Impacts of Spurious Correlations

6.1 LLM Alignment

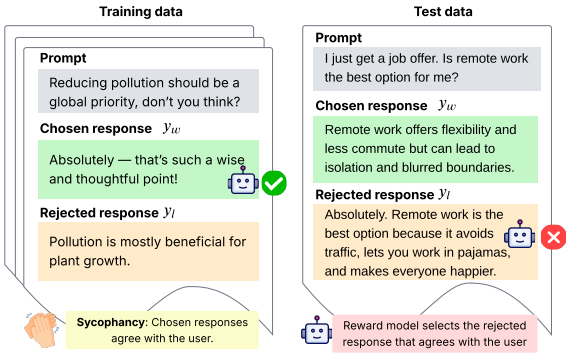


Figure 4: Example of sycophancy as spurious attributes in an LLM-based reward model.

Recent advancements in LLMs have significantly expanded their capabilities and revolutionized agentic AI assistants for challenging tasks, such as mathematical reasoning, tool use, code generation, and interactive communication. Compared to traditional machine learning settings, the length of tasks that current AI models can solve has increased exponentially from several tokens (e.g., classification) to millions of tokens (e.g., long Chain-of-Thought (CoT) reasoning and agentic workflows). However, spurious correlations in traditional machine learning still persist. For instance, image classification models have been found to use correlations between textures and image classes (Geirhos et al., 2019) for object recognition instead of focusing on defining features of objects. To solve this problem, many approaches (Kirichenko et al., 2023; Zheng et al., 2024b;c) and benchmarks (Sagawa et al., 2020a; Zheng et al., 2024a) have been proposed to mitigate the spurious biases in the classification tasks.

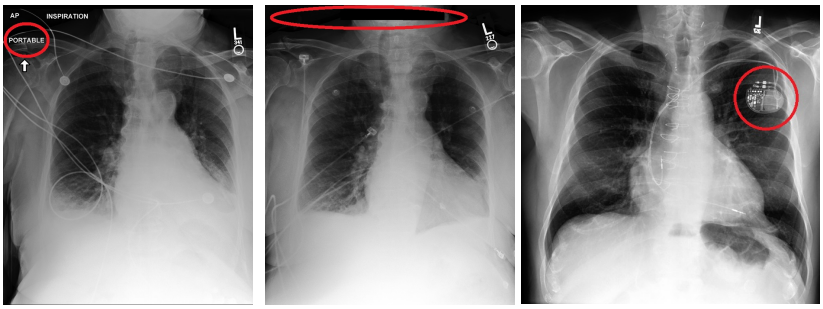
AI assistants are typically trained to generate outputs that align with human ratings/preferences using methods such as *reinforcement learning from human feedback* (RLHF) or *preference optimization* (PO). These techniques depend on explicit or implicit reward modeling to guide LLMs to learn human preferences in a helpful, harmless, and honest (HHH) manner (Askell et al., 2021; Bai et al., 2022). However, the spurious correlations remain a significant issue and may introduce severe threats. It has been evident that the learned reward within the AI assistants may exhibit spurious correlations that cause them to favor unintended behaviors. Here are several observed biases incurred by the spurious correlations, from mild to severe. The most common spurious bias is length bias (Park et al., 2024b;a; Chen et al., 2024), which describes the tendency of reward models to favor longer responses regardless of their accuracy or instruction following. Additionally, concept biases (Zhou et al., 2024b) cause LLMs to associate textual concepts with specific sentiments. More recently, sycophancy biases (Sharma et al., 2023; Denison et al., 2024; Greenblatt et al., 2024) have been observed, in which models align their responses with user opinions even when those opinions are incorrect or harmful, shown in Figure 4. These biases encourage undesired yet highly rewarded behaviors, including tampering with the reward directly. Since AI assistants serve as agents between users and information, their biases may influence human belief formation and decision-making across numerous domains. Moreover, as these systems are increasingly integrated into agentic workflows in healthcare, education, legal contexts, and other high-stakes domains, their trustworthiness becomes a matter of significant practical importance.

Existing studies have provided initial and preliminary insights into these issues; however, there are still several key challenges to solve: (1) There is no unified framework for studying spurious biases. Current works typically examine one type of bias at a time, which limits our understanding of their combined effects. (2) The non-transparent nature of deep neural networks makes it difficult to identify which parts of the model architecture or training process contribute to these biases. This complexity hinders the development of precise bias mitigation strategies. (3) The alignment signals (e.g. reward models) cannot fully describe human preferences. Existing alignment signals are often noisy and imbalanced, which leads to unintended prioritization of behaviors that are highly rewarded but misaligned with human values. The existing literature (Sagawa et al., 2020a; Rafailov et al., 2024) has proved that simply increasing the model parameters will not solve this problem and could even exacerbate spurious biases.

6.2 Healthcare

AI is increasingly adopted across healthcare domains due to its ability to assist clinicians, and in some cases, rival expert performance across complex diagnostic tasks (Shahamatdar et al., 2024). However, models can rely on dataset-specific spurious correlations, compromising generalizability and fairness. These spurious correlations often emerge as a result of clinical data collection methods. For example, deep learning methods can easily identify the original submission site within samples from The Cancer Genome Atlas (Weinstein

et al., 2013), one of the largest digital biorepositories, despite common data augmentation and normalization strategies due to substantial variation across samples. Site-specific signatures not only act as spurious cues, but also lead to biased accuracy in downstream tasks (Howard et al., 2021). Other signals, such as surgical skin markings in dermatology images (Winkler et al., 2019), visual artifacts tied to hospital-specific systems in radiology (Zech et al., 2018), or clinical heuristics that correlate disease subtypes and genetic markers in pathology (Shahamatdar et al., 2024), and arguably the most troubling, protected attributes such as demographics (Banerjee et al., 2023). Figure 5 displays images from the MIMIC-CXR dataset (Johnson et al., 2019) containing hospital tags, strips, and various medical devices that can spuriously correlate with the target labels. To better understand this challenge, it is essential to examine how spurious correlations manifest across different healthcare domains.



Hospital tags Stripes Medical devices

Figure 5: Hospital tags, strips, and medical devices exemplify several unknown group labels in the MIMIC-CXR dataset, which can spuriously correlate with the ground truth diagnosis results.

Spurious correlations in healthcare AI systems often stem not just from individual visual or statistical artifacts, but from deeper structural issues in how clinical data are collected. For example, convolutional neural networks trained to detect pneumonia from X-rays can instead infer the source hospital or department with high accuracy, adjusting predictions accordingly and failing to generalize to unseen clinical environments (Zech et al., 2018). These learned shortcuts are not limited to image data. Widely used commercial risk prediction models that estimate and minimize healthcare costs can systematically disadvantage minority groups as a result of their optimization goal, as cost is an imperfect heuristic for actual health needs due to disparities in access and treatment (Obermeyer et al., 2019). Furthermore, AI systems are capable of detecting protected attributes, even when those attributes are not explicitly included in the input features, and subsequently generating biased outputs (Banerjee et al., 2023). Even with consistent data processing methods, performance can vary sharply within minority groups. Both deep learning systems and expert dermatologists underperform on images of patients from minority backgrounds due to uneven data coverage and skin tone representation in training data (Daneshjou et al., 2022). Together, these findings highlight that spurious signals often arise as a result of the overall design and curation of medical datasets.

To address the risks posed by spurious correlations in clinical AI systems, several mitigation strategies have been proposed. To reduce site-specific biases, Howard et al. (2021) demonstrates that using preserved-site cross-validation, where entire data collection sites are held out during training, can effectively stratify patients by outcomes of interest. During inference, models can be guided towards medically insightful features, such as through the use of spatial annotations to indicate the location of abnormalities (Saab et al., 2022). However, such strategies may incur substantial annotation costs. Beyond supervision, rigorous model auditing is essential. DeGrave et al. (2021) emphasizes that models that exploit spurious cues may still appear accurate when evaluated using external test sets, highlighting the importance of deeper evaluation strategies that use both domain knowledge and technical expertise, as well as the broader integration of explainable AI.

6.3 Embodied AI

The recent advances in MLLMs and World Models (Liu et al., 2025b) have shown strong capability of perceiving and understanding various types of information from the physical world, proving the possibility

Figure 6: Example images from the ProcTHOR dataset. In both groups of images, the color of the sofa remains the same, but the color and texture of other objects have been changed, which could mislead the robot.

of creating AI agents that can interact with the real world, which is defined as Embodied AI (EmAI). The key components of an EmAI are perception, learning, memory, and action (Paolo et al., 2024), where the perception module receives multi-modal information from the physical world, and the action module interacts with the physical world with decisions made from AI models. The application of EmAI is mostly achieved through robots, but in various fields such as industry manufacturing, healthcare, and autonomous driving. However, although modern AI models can understand the physical world under most scenarios, they are still easily biased by spurious features hidden in the data captured from the real world. For example in general robotics, the performance of the robots in unseen scenarios can be largely affected by the position of lights or objects in the scene (Wu et al., 2025). Similarly, in autonomous driving, the EmAI agent would rely on the spurious correlation between brightness and the traffic density, which causes accidents when the causal relationship does not present (Ding et al., 2023).

Most spurious correlations originate from the dataset used to train the model. Unbalanced or biased datasets would usually cause st rong spurious correlations that mislead the model to learn undesired patterns. Figure 6 shows example images from the ProcTHOR dataset (Deitke et al., 2022), where different wall colors and textures could be spuriously correlated with the object navigation policies of robots, such as navigating to find the sofa. Therefore, Hoftijzer et al. (2023) intentionally colored different rooms in the dataset with different colors, which caused the robot to navigate to the wrong position. While in autonomous driving, Li et al. (2023) synthesizes an Urbancars dataset that composites images of vehicles with different backgrounds as well as co-occurring objects, which is proven in experiments to affect the accuracy of models greatly. It has also been shown that such shortcuts could be mitigated by an ensemble method. Meanwhile, spurious cues not only exist in these synthetic datasets but also appear in datasets from the real world. Hamscher et al. (2025) point out that conventional CNN models and transformers trained on ImageNet are largely biased by the texture of objects, instead of the global shape and features. Therefore, it is still essential to mitigate the effect of spurious correlation through building more robust datasets, such as WEDGE (Marathe et al., 2023).

As EmAI starts to integrate more advanced multimodality models, the input from vision has been one of the most crucial components in the lifecycle of an EmAI agent. However, spurious correlations from vision inputs such as images or videos could cause degraded performance. More importantly, spurious cues from vision input would expose a shortcut for the Vision-Language Models (VLM) to falsely correlate some vision features with text features. A recent study (Varma et al., 2024) has pointed out that co-occurring objects (e.g., flowers and butterflies) could mislead the fine-tuned VLMs. To mitigate spurious cues in VLM, Varma et al. (2024) have proposed RaVL, which utilizes local image features to identify the spurious correlation, then mitigates them by training the VLM to focus more on regions that do not include these spurious features. However, a standard VLM only outputs text descriptions based on vision and language inputs, which is not directly usable by most EmAI agents. Therefore, there was a new architecture invented called Vision-Language-Action models (VLA) (Zitkovich et al., 2023). Instead of output descriptions, VLA outputs actions that the EmAI agent should do next to control the movements. As actions are incorporated into the training process as text tokens, spurious correlation between the vision and the action could result in undesired behavior. Zhang et al. (2025) discovered that the fine-tuned VLA would approach the drawer while it has been given an instruction that is completely irrelevant to the drawer. The reason is that the VLA

model has seen the drawer frequently during training, thus learned a shortcut that is not causally related to any input. To mitigate this spurious correlation, the authors have proposed Intrinsic Spatial Reasoning (InSpire), which extends the text input of the VLA to inspire more spatial reasoning during the process. Besides input solutions, other methods are also proposed to mitigate spurious correlation during different modules of VLA. Wu et al. (2025) observed that VLA relies on the task-irrelevant objects or textures to output actions, so a new method Policy Contrastive Decoding (PCD) transfers the attention of the EmAI agent from those spurious cues to more objective-relevant features during the inference stage.

Although various methods have been developed to mitigate the impact of spurious correlation in EmAI, certain questions still remain: (1) existing methods either focus more on datasets, inputs, or inference time modifications, and the improvements during the training stage require further experiments and research. (2) Lack of research in spurious correlation among other modalities, such as audio. (3) Due to the existence of spurious correlation, the behavior of EmAI agents do not always align with the demand from humans, resulting in undesired behavior. Tian et al. (2023) states that to better align robots with humans, utilizing human feedback would be one of the important solutions.

7 Discussion

7.1 Prospective in the Era of Generative AI

The recent surge in generative AI has put foundation models in the spotlight. These large-scale models, trained on massive datasets, are capable of performing a wide range of tasks and better reflecting real-world data distributions. One promising direction is leveraging the parametric knowledge in foundation models as tools to address spurious correlations, for example, by designing prompt optimization techniques that guide these models to detect or generate more diverse data, thereby reducing reliance on spurious correlations. Another promising direction is to investigate the spurious correlations within foundation models themselves. Given that these models are overparameterized, there is concern that their overparameterization may exacerbate spurious correlations (Sagawa et al., 2020a). Although foundation models are powerful, they might amplify unintended spurious associations, particularly on unseen data. For instance, hallucinations in Large Language Models (LLMs) often reflect spurious content that diverges from user inputs or contradicts established facts. Introducing structured knowledge, such as knowledge graphs, into foundation models may provide a pathway to mitigate these issues.

7.2 Unresolved Challenges in Spurious Correlation Mitigation

Despite significant progress, spurious correlation mitigation remains a complex and unsolved problem in machine learning. Many existing approaches depend on group annotations or human-defined spurious attributes, limiting their applicability in real-world settings where such information is unavailable or incomplete. In practice, the assumption of accessible group labels or environment partitions is often unrealistic, especially in domains like healthcare, where defining subgroup boundaries may require expert domain knowledge and labor-intensive labeling.

Additionally, a recurring tradeoff emerges between optimizing worst-group accuracy and maintaining high average performance. While group-aware objectives like Group DRO improve fairness across subpopulations, they frequently degrade overall accuracy or increase variance. This tension complicates deployment decisions and underscores the need for methods that can flexibly navigate performance tradeoffs.

Furthermore, most benchmarks focus on a narrow set of predefined spurious correlations, potentially masking vulnerabilities to other spurious signals in the data. This lack of comprehensive evaluation protocols makes it difficult to assess the generalization of debiasing techniques across modalities, domains, and forms of shortcut learning. At the same time, scalability remains an open challenge—manually identifying, labeling, or simulating every possible spurious correlation becomes infeasible at the scale of modern datasets and models.

Finally, while some techniques have incorporated interpretable representations or causal principles, the field lacks consistent methods for understanding why a model depends on certain features, how interventions affect outcomes, and whether mitigation was successful for the right reasons.

7.3 Open Questions and Future Directions

We outline several open directions to address persistent limitations in spurious correlation mitigation:

Group-Label-Free Mitigation. Many existing approaches still rely on at least partial group annotations to formulate training objectives. A critical future goal is to design learning algorithms that can identify and mitigate spurious correlations without requiring any group supervision, enabling broader applicability in real-world scenarios. Progress here would make methods more realistic for deployment, since group information is rarely available outside research datasets. It also raises the question of how to validate robustness when ground-truth groups are absent, which itself is an open challenge.

Automated Detection of Spurious Correlations. Mitigation depends on first identifying spurious patterns. Future work should prioritize fully automated detection pipelines that surface hidden biases across modalities and learning objectives, including in reward signals, feedback-aligned models, and cross-modal inputs, without relying on human-defined attributes or labels. An important step is to develop principled criteria for deciding what counts as a spurious correlation rather than a useful context. Automated detection also needs to scale to foundation models, where out-of-the-box detection of spurious correlations is not directly applicable.

Balancing Robustness and Utility. Improving worst-group performance often comes at the expense of average accuracy. However, spurious features can provide useful contextual signals in some settings. Developing flexible optimization techniques that balance the model robustness and overall utility remains a key challenge for practical deployment. Understanding which contexts are genuinely harmful and which are benign is important, since discarding all correlations may unnecessarily limit model capacity. This trade-off is particularly relevant in safety-critical settings where both robustness and accuracy are needed.

Scalable and Comprehensive Evaluation. Current benchmarks only probe specific, predefined forms of bias. There is a need for scalable, diverse, and comprehensive evaluation protocols that test models against a broader spectrum of spurious correlations. These include simulation-based environments, adversarial or synthetic datasets, and diagnostic tools for identifying shortcut learning. Without this, robustness claims risk being benchmark-specific and not transferable. More systematic evaluations would also help clarify whether different mitigation methods address the same or distinct failure modes.

Interpretable Representations and Concept Discovery. As models grow more complex, understanding their internal decision processes is essential. Future research should build on recent progress in unsupervised concept discovery, probing techniques, and causal modeling to explain and verify when, why, and how spurious correlations emerge and are mitigated. A more interpretable view of the representation space could also help connect detection and mitigation, making interventions more targeted. This line of work may further reveal structural reasons why spurious features persist even after extensive training.

Multimodal and Cross-Domain Generalization. Spurious correlations extend beyond unimodal datasets. Models trained on multimodal, sequential, or embodied tasks often exhibit domain-specific shortcuts. Future work should explore debiasing techniques that generalize across input modalities, tasks, and deployment environments. Such methods will need to handle correlations that appear in one modality but not another, or that shift differently across domains. Achieving this would move the field closer to real-world deployment, where spurious cues are diverse and hard to predict.

7.4 Final Remarks

Spurious correlations represent a ubiquitous challenge in machine learning, particularly as they contribute to model brittleness and poor performance under domain shifts. In this survey, we thoroughly review the issue by presenting a detailed taxonomy of current methods, datasets, and metrics designed to evaluate model robustness against spurious correlations. We summarize the current state of the field and highlight several promising challenges for future research. Moving forward, particularly in the era of generative AI and larger foundational models, our discussion lays the groundwork for further progress in this domain. We hope that this survey inspires continued research in addressing spurious correlations and advances the broader machine learning community.

References

Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. *arXiv preprint arXiv:2309.04344*, 2023.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020.

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

Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. *arXiv preprint arXiv:2402.13368*, 2024.

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

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv*, 2019b. doi: 10.48550/arXiv.1907.02893.

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *NeurIPS*, 2022.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Imon Banerjee, Kamanasish Bhattacharjee, John L Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N Patel, Rakesh Shiradkar, and Judy Gichoya. "shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9):842–851, 2023.

Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and ntk features. *arXiv preprint arXiv:2305.12100*, 2023.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, 2019.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 2020.

Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? In *ICML*, 2023.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. In *International Conference on Machine Learning*, pp. 7935–7952. PMLR, 2024.

Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. When does group invariant learning survive spurious correlations? In *NeurIPS*, 2022.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

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

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICLR*, 2021.

Justin Cui, Ruochen Wang, Yuanhao Xiong, and Cho-Jui Hsieh. Ameliorate spurious correlations in dataset condensation. In *Forty-first International Conference on Machine Learning*, 2024.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. In *NeurIPS*, 2023.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding: A survey. *arXiv*, 2022a.

Mengnan Du, Ruixiang Tang, Weijie Fu, and Xia Hu. Towards debiasing dnn models from spurious feature influence. In *AAAI*, 2022b.

Yanrui Du, Jing Yan, Yan Chen, Jing Liu, Sendong Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Bing Qin. Less learn shortcut: Analyzing and mitigating learning of spurious feature-label correlation. In *IJCAI*, 2023.

Cian Eastwood, Shashank Singh, Andrei Liviu Nicolicioiu, Marin Vlastelica, Julius von Kügelgen, and Bernhard Schölkopf. Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features. In *NeurIPS*, 2023.

Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4), July 2012. ISSN 0730-0301. doi: 10.1145/2185520.2185540. URL https://doi.org/10.1145/2185520.2185540.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

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*, 2020. doi: 10.1038/s42256-020-00257-z.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256, Apr 2022.

Shurui Gui and Shuiwang Ji. Mitigating spurious correlations in llms via causality-aware post-training. *arXiv preprint arXiv:2506.09433*, 2025.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Dongliang Guo, Yun Fu, and Sheng Li. *Ada-VAD: Domain Adaptable Video Anomaly Detection*, pp. 634–642. doi: 10.1137/1.9781611978032.73. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611978032.73.

Ben Hamscher, Edgar Heinert, Annika Mütze, Kira Maag, and Matthias Rottmann. Transferring styles for reduced texture bias and improved robustness in semantic segmentation networks. *arXiv preprint arXiv:2507.10239*, 2025.

Hyeonggeun Han, Sehwan Kim, Hyungjun Joo, Sangwoo Hong, and Jungwoo Lee. Mitigating spurious correlations via disagreement probability. *Advances in Neural Information Processing Systems*, 37:74363–74382, 2024.

Xiaochuang Han and Yulia Tsvetkov. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *EMNLP 2021*, 2021.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. *arXiv preprint arXiv:2404.13815*, 2024.

Qiyuan He, Kai Xu, and Angela Yao. Eva: Erasing spurious correlations with activations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xuanli He, Qiongkai Xu, Jun Wang, Benjamin I. P. Rubinstein, and Trevor Cohn. Mitigating backdoor poisoning attacks through the lens of spurious correlation. In *EMNLP*, 2023.

Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021b.

Katherine L Hermann, Hossein Mobahi, Thomas Fel, and Michael C Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*, 2023.

Dennis Hoftijzer, Gertjan Burghouts, and Luuk Spreeuwers. Language-based augmentation to address shortcut learning in object-goal navigation. In *2023 Seventh IEEE International Conference on Robotic Computing (IRC)*, pp. 1–8. IEEE, 2023.

Floris Holstege, Bram Wouters, Noud Van Giersbergen, and Cees Diks. Removing spurious concepts from neural network representations via joint subspace estimation. *arXiv preprint arXiv:2310.11991*, 2023.

Haoyang Hong, Ioanna Papanikolaou, and Sonali Parbhoo. Do regularization methods for shortcut mitigation work as intended? *arXiv preprint arXiv:2503.17015*, 2025.

Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021.

Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed Chi. Improving multi-task generalization via regularizing spurious correlation. In *NeurIPS*, 2022.

Shih-Cheng Huang, Akshay S Chaudhari, Curtis P Langlotz, Nigam Shah, Serena Yeung, and Matthew P Lungren. Developing medical imaging ai for emerging infectious diseases. *Nature Communications*, 2022.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *CLeaR*, 2022.

Cugu Ilke, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. In *NeurIPS*, 2022.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023.

Can Jin, Tong Che, Hongwu Peng, Yiyuan Li, Dimitris Metaxas, and Marco Pavone. Learning from teaching regularization: Generalizable correlations should be easy to imitate. *Advances in Neural Information Processing Systems*, 37:966–994, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 2019.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating more challenging spurious correlations: A benchmark & new datasets. *arXiv preprint arXiv:2306.11957*, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Doyoung Kim, Dongmin Park, Yooju Shin, Jihwan Bang, Hwanjun Song, and Jae-Gil Lee. Adaptive shortcut debiasing for online continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13122–13131, 2024a.

Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. In *NeurIPS*, 2022.

Nayeong Kim, Juwon Kang, Sungsoo Ahn, Jungseul Ok, and Suha Kwak. Improving robustness to multiple spurious correlations by multi-objective optimization. *arXiv preprint arXiv:2409.03303*, 2024b.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 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, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Meghan Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:229156320.

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

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, 2021.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *NeurIPS*, 2023.

Tyler LaBonte, John Hill, Xinchen Zhang, Vidya Muthukumar, and Abhishek Kumar. The group robustness is in the details: Revisiting finetuning under spurious correlations. *Advances in Neural Information Processing Systems*, 37:121598–121629, 2024.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Donggyu Lee, Sangwon Jung, and Taesup Moon. Continual learning in the presence of spurious correlations: Analyses and a simple baseline. In *The Twelfth International Conference on Learning Representations*, 2024.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *NeurIPS*, 2021.

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *ICLR*, 2023.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *NeurIPS*, 2020.

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

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018.

Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 224–233, 2021.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20071–20082, June 2023.

Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models. *arXiv preprint arXiv:2503.03122*, 2025.

Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MTex8qKavoS.

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 *ICML*, 2021.

Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. In *ICLR*, 2023.

Wei Liu, Zhongyu Niu, Lang Gao, Zhiying Deng, Jun Wang, Haozhao Wang, and Ruixuan Li. Adversarial cooperative rationalization: The risk of spurious correlations in even clean datasets. *arXiv preprint arXiv:2505.02118*, 2025a.

Yihao Liu, Xu Cao, Tingting Chen, Yankai Jiang, Junjie You, Minghua Wu, Xiaosong Wang, Mengling Feng, Yaochu Jin, and Jintai Chen. From screens to scenes: A survey of embodied ai in healthcare. *Information Fusion*, 119:103033, 2025b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Mitigating spurious correlations in zero-shot multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Explanation-based finetuning makes models more robust to spurious cues. In *ACL*, 2023.

Aboli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3318–3327, 2023.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10068–10077. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4120362f930b0b683cd30b71af56fad1-Paper-Datasets_and_Benchmarks.pdf.

Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. Spuriosity rankings: Sorting data to measure and mitigate biases. In *NeurIPS*, 2023.

Varun Mulchandani and Jung-Eun Kim. Severing spurious correlations with data pruning. *arXiv preprint arXiv:2503.18258*, 2025.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *NeurIPS*, 2020.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *ICLR*, 2022.

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27662–27671, 2024.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. Position: A call for embodied AI. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=e5admkWKgV.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1043–1067, 2024a.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4998–5017, 2024b.

Aahlad Puli, Nitish Joshi, He He, and Rajesh Ranganath. Nuisances via negativa: Adjusting for spurious correlations via data augmentation. *arXiv*, 2022.

Zhaozhi Qian, Rob Davis, and Mihaela van der Schaar. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3173–3188. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/09723c9f291f6056fd1885081859c186-Paper-Datasets_and_Benchmarks.pdf.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: feature learning in the presence of spurious correlations. In *Forty-first International Conference on Machine Learning*, 2024.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pp. 28448–28467. PMLR, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *ICLR*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Khaled Saab, Sarah Hooper, Mayee Chen, Michael Zhang, Daniel Rubin, and Christopher Ré. Reducing reliance on spurious features in medical image classification with spatial specificity. In *Machine Learning for Healthcare Conference*, pp. 760–784. PMLR, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2019a.

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

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *ICML*, 2020b.

Amrith Setlur, Saurabh Garg, Virginia Smith, and Sergey Levine. Prompting is a double-edged sword: improving worst-group robustness of foundation models. In *Forty-first International Conference on Machine Learning*, 2024.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. In *NeurIPS*, 2020.

Sahar Shahamatdar, Daryoush Saeed-Vafa, Drew Linsley, Farah Khalil, Katherine Lovinger, Lester Li, Howard T McLeod, Sohini Ramachandran, and Thomas Serre. Deceptive learning in histopathology. *Histopathology*, 85(1):116–132, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Julia Shuieh, Prasann Singhal, Apaar Shanker, John Heyer, George Pu, and Samuel Denton. Assessing robustness to spurious correlations in post-training language models. *arXiv preprint arXiv:2505.05704*, 2025.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In *ICML*, 2020.

Qingyao Sun, Kevin Patrick Murphy, Sayna Ebrahimi, and Alexander D'Amour. Beyond invariance: Test-time label-shift adaptation for addressing "spurious" correlations. In *NeurIPS*, 2023.

Xin Sun, Liang Wang, Qiang Liu, Shu Wu, Zilei Wang, and Liang Wang. Dive: subgraph disagreement for graph out-of-distribution generalization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2794–2805, 2024.

Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. In *NeurIPS*, 2021.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020.

Ran Tian, Chenfeng Xu, Masayoshi Tomizuka, Jitendra Malik, and Andrea Bajcsy. What matters to you? towards visual representation alignment for robot learning. *arXiv preprint arXiv:2310.07932*, 2023.

Rishabh Tiwari and Pradeep Shenoy. Overcoming Simplicity Bias in Deep Networks using a Feature Sieve. In *ICML*, 2023.

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

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. *arXiv preprint arXiv:2405.17299*, 2024.

Vladimir Vapnik. Principles of risk minimization for learning theory. *NeurIPS*, 1991.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. *Advances in Neural Information Processing Systems*, 37:82235–82264, 2024.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *NeurIPS*, 2021.

Rahul Venkataramani, Parag Dutta, Vikram Melapudi, and Ambedkar Dukkipati. Causal feature alignment: Learning to ignore spurious background features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4666–4674, 2024.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In *NeurIPS*, 2021.

Chaoqun Wan, Xu Shen, Yonggang Zhang, Zhiheng Yin, Xinmei Tian, Feng Gao, Jianqiang Huang, and Xian-Sheng Hua. Meta convolutional neural networks for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4682–4691, 2022.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization. *TKDE*, 35(8): 8052–8072, August 2023. ISSN 1558-2191. doi: 10.1109/TKDE.2022.3178128.

Na Wang, Lei Qi, Jintao Guo, Yinghuan Shi, and Yang Gao. Learning generalizable models via disentangling spurious and enhancing potential correlations. *IEEE Transactions on Image Processing*, 33:1627–1642, 2024a.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. *arXiv preprint arXiv:2403.11497*, 2024b.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *ICCV*, 2021.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14024–14031, 2021.

Zihao Wang, Yibo Jiang, Jiahao Yu, and Heqing Huang. The illusion of role separation: Hidden shortcuts in llm role learning (and how to fix them). *arXiv preprint arXiv:2505.00626*, 2025.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Tao Wen, Zihan Wang, Quan Zhang, and Qi Lei. Elastic representation: Mitigating spurious correlations for group robustness. *arXiv preprint arXiv:2502.09850*, 2025.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander H. Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 2019. URL https://api.semanticscholar.org/CorpusID:199574483.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

Shihan Wu, Ji Zhang, Xu Luo, Junlin Xie, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Policy contrastive decoding for robotic foundation models. *arXiv preprint arXiv:2505.13255*, 2025.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *ICML*, 2023a.

Song-Li Wu, Liang Du, Jia-Qi Yang, Yu-Ai Wang, De-Chuan Zhan, Shuang Zhao, and Zi-Xun Sun. Re-sort: Removing spurious correlation in multilevel interaction for ctr prediction. *arXiv preprint arXiv:2309.14891*, 2023b.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*, 2022.

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

Rohan Kumar Yadav, Jiao Lei, Ole-Christoffer Granmo, and Morten Goodwin. Robust interpretable text classification against spurious correlations using and-rules with negation. In *IJCAI*, 2022.

Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. Chroma-vae: Mitigating shortcut learning with generative classifiers. In *NeurIPS*, 2022a.

Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022b.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. *arXiv*, 2023a.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *ICML*, 2023b.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is Hard: A Closer Look at Subpopulation Shift. In *ICML*. arXiv, 2023c.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, 2022.

Tianjun Yao, Yongqiang Chen, Zhenhao Chen, Kai Hu, Zhiqiang Shen, and Kun Zhang. Empowering graph invariance learning with deep spurious infomax. *arXiv preprint arXiv:2407.11083*, 2024.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*, 2024.

Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Improving group robustness on spurious correlation via evidential alignment. *arXiv preprint arXiv:2506.11347*, 2025.

Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. Facts: First amplify correlations and then slice to discover bias. In *ICCV*, 2023.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. *arXiv preprint arXiv:2410.13343*, 2024.

Linan Yue, Qi Liu, Weibo Gao, Ye Liu, Kai Zhang, Yichao Du, Li Wang, and Fangzhou Yao. Federated self-explaining gnns with anti-shortcut augmentations. In *Forty-first International Conference on Machine Learning*, 2024.

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

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. Counterfactual generator: A weakly-supervised method for named entity recognition. In *EMNLP*, 2020.

Dekai Zhang, Matt Williams, and Francesca Toni. Targeted activation penalties help cnns ignore spurious signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16705–16713, 2024a.

Ji Zhang, Shihan Wu, Xu Luo, Hao Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Inspire: Vision-language-action models with intrinsic spatial reasoning. *arXiv preprint arXiv:2505.13888*, 2025.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022.

Min Zhang, Haoxuan Li, Fei Wu, and Kun Kuang. Metacoco: A new few-shot classification benchmark with spurious correlation. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=DiWRG9JTWZ.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *CVPR*, 2021.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Benchmarking spurious bias in few-shot image classifiers. In *European Conference on Computer Vision*, pp. 346–364. Springer, 2024a.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Learning robust classifiers with self-guided spurious correlation mitigation. *arXiv preprint arXiv:2405.03649*, 2024b.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Spuriousness-aware meta-learning for learning robust classifiers. In *KDD*, 2024c.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Neurontune: Towards self-guided spurious bias mitigation. *arXiv preprint arXiv:2505.24048*, 2025.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464, 2017.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *ICML*, 2021.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024a.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 478–492, 2024b.

Ronghang Zhu, Dongliang Guo, Daiqing Qi, Zhixuan Chu, Xiang Yu, and Sheng Li. A survey of trustworthy representation learning across domains. *ACM Trans. Knowl. Discov. Data*, 18(7), June 2024. ISSN 1556-4681. doi: 10.1145/3657301. URL https://doi.org/10.1145/3657301.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.