Concept-Residual Representation Expansion for Robustness to Spurious Correlations

Eric Xie
University of Virginia
Charlottesville, USA
jrg4wx@virginia.edu

Guangzhi Xiong
University of Virginia
Charlottesville, USA
hhu4zu@virginia.edu

Wenqian Ye
University of Virginia
Charlottesville, USA
wenqian@virginia.edu

Aidong Zhang
University of Virginia
Charlottesville, USA
aidong@virginia.edu

Abstract

Models trained with empirical risk minimization (ERM) are prone to relying on spurious correlations to make predictions. A spurious correlation is a non-causal relationship in the training data between an attribute and the prediction target that does not generalize beyond the training environment. As a result, models can appear to achieve strong performance by exploiting these correlations, yet fail when the correlation changes or disappears. Despite their tendency to learn spurious correlations, the success of post-hoc mitigation methods in recent work suggests that ERM-trained models still retain useful, robust predictive features. However, core (non-spurious) features may be weak or entangled within the representation, making them difficult to identify. We propose Concept-Residual eXpansion (CRX), a concept-augmented framework that improves robustness by expanding the set of candidate predictive features. Starting from a frozen ERM representation, we augment the model's features with interpretable concept scores that describe the presence of task-relevant attributes and the surrounding context, together with residual features that capture the portion of the ERM features not expressed by the concepts. We then retrain a lightweight classifier on this expanded feature space, enabling it to leverage both structured semantic cues and complementary residual information. Across standard spurious correlation benchmarks, CRX consistently improves worst-group accuracy while maintaining competitive average performance. These results suggest that expanding the set of available features can substantially improve robustness to spurious correlations. The code and additional implementation details can be found at https://doi.org/10.5281/zenodo.20467988

CCS Concepts

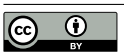
• **Computing methodologies** → **Machine learning**; *Neural networks*; *Computer vision*; **Learning latent representations**.

Keywords

Spurious Correlations; Group Robustness; Worst-Group Accuracy; Representation Learning; Concept-Based Models; Vision-Language Models

ACM Reference Format:

Eric Xie, Guangzhi Xiong, Wenqian Ye, and Aidong Zhang. 2026. Concept-Residual Representation Expansion for Robustness to Spurious Correlations. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3770855.3818089



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818089

1 Introduction

A spurious correlation refers to a non-causal relationship between a prediction target and an attribute that is not essential for defining the target, commonly referred to as a spurious attribute. Training data often contains such correlations that do not persist outside the training environment, leading to models with high average performance that severely fail in specific subpopulations [13, 43]. Addressing this issue is a central focus of group robustness methods. In classification tasks, a group label (class, spurious attribute) annotates a sample with a spurious attribute and its class label, representing a spurious correlation. For example, the prediction target "landbird" may become spuriously associated with a land background due to frequent co-occurrence in pre-training data. As a result, a model relying on spurious correlations may incorrectly predict a waterbird as a landbird simply because it appears on a land background.

Empirical risk minimization (ERM) [32] learns representations that capture both predictive structure and correlations specific to the training environment. Failures can arise when predictors rely heavily on the latter [31]. For example, models intended to identify the presence of pneumonia were shown to instead rely on spurious correlations between metal tokens on chest radiographs from different hospitals and disease detection results [13, 41], rather than the pathological features of pneumonia itself. Recent studies find that strong performance on underperforming groups through sample reweighting [10, 23], distributionally robust optimization (DRO) [28], or a post-hoc classifier [18, 20] can substantially reduce reliance on such spurious correlations. Since these methods primarily reweight or learn from existing features rather than introducing new information, they suggest that useful features for robust prediction are already present in the learned representation within ERM models in many cases [17, 39]. This highlights the potential of using the model's internal representation to improve robustness.

However, such mitigation methods have inherent limitations in practical applications. In real-world settings, group labels such as (landbird, land) or (waterbird, water) are rarely available, making strategies that rely on group annotations difficult to apply [11, 40]. Moreover, even when robust features exist in principle, they may not be encoded in the latent space in a way that is easily separable or usable by downstream predictors. Subtle attributes or group-specific patterns can be weakly represented or entangled with contextual correlations [6]. In such cases, improving how we select predictors

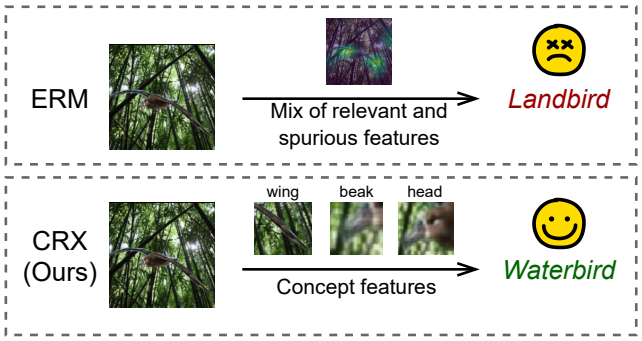


Figure 1: Illustrative comparison between a standard ERM classifier and the proposed Concept-Residual eXpansion (CRX) framework on an image of a waterbird.

may not be sufficient if the underlying representation does not provide strong alternatives.

Instead, in this work, we aim to enrich the representation itself by expanding it with additional semantically meaningful features. We introduce a concept-centric framework that augments ERM features with two additional sources of structured information. First, we construct a vocabulary of visual concepts using a language model and estimate their presence in images using CLIP [27]. These concept features expose interpretable attributes related to object structure, context, and co-occurring elements. Second, because a finite concept set cannot capture all task-relevant variation, we introduce residual features that model the portion of the ERM representation not explained by the concept space. Together, these components provide a more structured feature space for downstream learning.

Figure 1 illustrates the comparison between a standard ERM classifier and our proposed method on a waterbird image. The ERM model entangles label-relevant cues with spurious background features, leading to incorrect "landbird" predictions when the waterbird appears on land. Instead, our method augments the ERM representation with concept scores (e.g., wing, beak, head) to emphasize object-centric evidence and recover the correct "waterbird" label, improving robustness to spurious correlations.

Our framework is designed to enable robust feature learning without requiring group labels at any stage of training or model selection. By expanding the representation with concept and residual features, we provide the model with a structured set of candidate features from which to form predictions. This enables robustness gains even when the spurious attributes are unknown or unavailable by encouraging the classifier to rely on stable predictive features rather than requiring explicit group supervision.

We evaluate this approach on standard spurious correlation benchmarks, without using any group labels during training or model selection. Our goal is to understand whether expanding the representation with concepts and residuals can improve the set of candidate features for downstream prediction. The results presented in this paper suggest that representation expansion offers an effective pathway to improving robustness under spurious correlations.

In summary, our contributions are:

- **A concept-based representation-expansion framework for robustness.** We introduce Concept-Residual eXpansion (CRX), which improves robustness by augmenting an ERM representation with structured semantic concept features and complementary residual features.
- **Group-free robustness gains through representation expansion.** Our empirical analysis demonstrates that CRX consistently improves worst-group accuracy on standard spurious correlation benchmarks without the use of group labels for training, validation, or model selection.
- **Empirical evidence that representation expansion drives robustness.** Through ablation studies, we show that augmenting ERM features with either concepts or residuals alone yields substantial improvements, while CLIP features without additional concept and residual components do not produce similar gains.

2 Related Works

Robustness to Spurious Correlations. Recent work on mitigating spurious correlations has shown that substantial robustness gains can be achieved by encouraging models to rely less on environment-specific cues and more on those that generalize across groups [20, 39]. Empirical risk minimization (ERM) implicitly assumes that training and test data are drawn from the same distribution [32]. When this assumption is violated, for example, when certain attributes co-occur with labels in training but not at test time, ERM can overemphasize spurious shortcuts that reduce average training loss but fail under spurious correlations. Many robust learning approaches address this issue by reweighting training examples, enforcing group-wise performance, or applying post-hoc adjustments, which can be interpreted as improving how predictors are selected from an ERM-learned representation [7, 15, 28]. Empirically, these methods are particularly effective when group labels are available during training or validation, suggesting that in many settings, the necessary information for robust prediction is already encoded but not optimally utilized.

However, these strategies often rely on access to group or environment labels, which are expensive to collect and rarely available in real-world applications [22, 24, 31]. This limits the practical applicability of methods that depend on explicit group supervision for training or model selection.

A growing line of work aims to improve robustness without relying on explicit group labels. Approaches include sample reweighting based on loss dynamics, confidence, or agreement across models; training procedures that separate invariant and variant features across environments; and post-hoc adjustments that retrain lightweight classifiers on frozen representations to emphasize more stable features [9, 12, 18, 24]. Although these strategies differ in how they estimate sample importance, each approximates the behavior of group-aware training without access to group labels. In effect, they attempt to identify which training examples correspond to underperforming or minority groups and increase their influence during learning. However, like group-supervised methods, they operate within the fixed representation learned by ERM. Their success therefore depends on the presence of sufficiently strong and separable causal features in the original latent space.

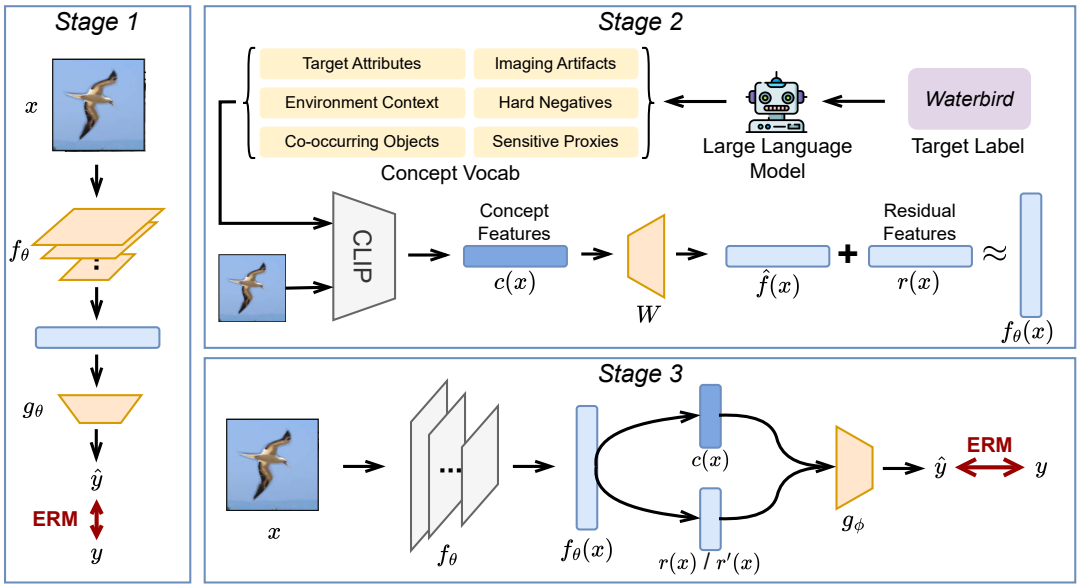


Figure 2: Schematic overview of the proposed Concept-Residual eXpansion (CRX) framework. First, an ERM-trained network produces a base representation f_θ. Next, a large language model uses the target labels to generate a concept vocabulary, to be scored using a CLIP model to obtain concept features $c(x)$ and derive residual features $r(x)$. A lightweight classifier g_ϕ is then trained on the expanded representation for prediction. Orange blocks denote trainable components, while blue blocks indicate intermediate computed features.

Beyond reweighting and post-hoc selection strategies, spurious correlations can also be mitigated by explicitly modifying the learned representation. Causal intervention methods attempt to model or remove confounding structure by identifying causal relationships between inputs, labels, and nuisance factors. Examples include approaches that incorporate causal attention modules [35] or semantic editing [3] to reduce confounding effects, or use structured causal graphs to model dependencies [30, 33]. Another major direction focuses on learning representations that are less sensitive to spurious variation. Invariant representation learning encourages features that remain stable among environments [5, 9, 34], while feature disentanglement methods separate shortcut-related and label-relevant components within the latent space through architectural design or subspace decomposition [37, 38, 42].

Together, prior work suggests that robustness can be improved either through better selection or explicit suppression of shortcut-related variation from a learned representation. However, many selection-based approaches are limited by the structure of the ERM feature space, while representation-focused methods often requires additional assumptions or supervision and risk removing stable, predictive features. This leaves an important middle ground. Rather than removing or enforcing invariance over specific features, the representation can instead be expanded with additional feature dimensions that make stable predictors easier to identify.

Concept-Based Representations. A complementary line of work seeks to improve robustness by introducing structured, semantically meaningful representations in addition to standard learned features. These approaches leverage concepts to provide models with axes of variation that are aligned with human-understandable and potentially causal factors [14, 19, 21].

Concepts can be used to correct spurious correlations, as interpretable concept axes can help expose shortcut features and guide debiasing. For example, identifying unstable or environment-dependent concepts and then intervening on the data or training objective can reduce reliance on spurious attributes [26, 36].

Concepts can also be used to construct structured feature spaces without requiring group annotations. Unsupervised or zero-shot concept discovery methods leverage object-centric learning or language models to define semantic subspaces that can improve robustness under spurious correlations [2, 4]. These approaches demonstrate that concept-aligned representations can mitigate spurious correlations even in bias-unsupervised settings, highlighting the value of semantically meaningful structure beyond standard latent features.

However, existing concept-based approaches typically use concepts as substitutes for learned representations, either by training concept bottlenecks [21], intervening on concept predictions [36], or reshaping embeddings through concept-aligned subspaces [19]. While this can improve interpretability and robustness, it risks discarding predictive cues that are difficult to describe in natural language or not included in the concept vocabulary.

Our work builds on this idea but takes a different perspective. Rather than replacing or intervening on learned features using concepts, we treat concepts as a structured augmentation to the original representation. We combine semantically meaningful concept features with residual features that capture information not explained by the concept space. This expanded representation increases the

set of candidate predictors available for robust selection, enabling improved worst-group performance even when group labels are unavailable during training or model selection.

3 Methodology

3.1 Overview

We propose a three-stage framework, displayed in Figure 2, that expands an ERM-learned representation with structured semantic features and complementary residual information, and then performs group-free prediction on the resulting feature space: (1) **ERM Representation Learning:** Train a standard model with ERM and freeze its penultimate-layer representation, (2) **Concept and Residual Feature Construction:** Construct an expanded feature space consisting of interpretable concept features and complementary residual features derived from ERM features after removing concept-explainable variation, and (3) **Expanded Feature Classification:** Train a lightweight classifier on the expanded representation with structured regularization, without using group or environment labels for training or model selection. We describe each stage in detail below.

3.2 ERM Representation Learning

We consider a supervised classification setting where each input image $x \in \mathcal{X}$ is associated with a label $y \in \mathcal{Y}$. In addition to label-relevant information, images may contain spurious factors that correlate with the label in the training data but do not reflect the true causal mechanism for prediction.

We begin by training a standard neural network using ERM on the original training data. Let $f_\theta(x) \in \mathbb{R}^d$ denote the penultimate-layer feature representation learned by this network with the learned parameters θ. The model is trained with a standard cross-entropy objective:

$$\min_{\theta} \ \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} \left[\ell_{\text{CE}}(g_\theta(f_\theta(x)), y)\right], \tag{1}$$

where g_θ is the final classifier and ℓ_{CE} is the cross-entropy loss.

We preserve the ERM representation and treat $f_\theta(x)$ as a high-capacity feature space that encodes a mixture of useful and misleading cues. The key challenge is then how to expose and disentangle this information so that downstream training can more easily select stable, predictive non-spurious features. To achieve this, we propose two complementary components from the ERM feature representation: an interpretable concept-based component and a residual component that captures any remaining predictive structure not covered by concepts.

3.3 Concept and Residual Feature Construction

While ERM representations encode rich visual information, their latent spaces are not structured to expose which dimensions correspond to stable, interpretable features versus spurious correlations. In this stage, we decompose the ERM feature representation into two complementary components: an interpretable concept-based representation and a residual representation that captures remaining predictive information.

Concept Bank Construction. We begin by constructing a task-specific vocabulary of visual concepts with a large language model (e.g., GPT-4o-mini), where the concepts can be scored from a single still image and organized into a small set of semantically meaningful channels. Rather than treating all concepts as a flat set of features, we group them into categories that reflect different sources of predictive information:

- Target Attributes (object-centric parts and morphology),
- Environment Context (scene and background cues),
- Co-Occurring Objects (objects frequently appearing nearby),
- Imaging Artifacts (capture and preprocessing artifacts),
- Hard Negatives (visually confusable categories),
- Sensitive Proxies (diagnostic-only concepts that may correlate with sensitive attributes).

This channel structure accomplishes two tasks. First, it encourages the concept vocabulary to cover multiple types of visual evidence rather than focusing narrowly on task-specific attributes. Second, they provide a mechanism for analysis and control within a task, allowing categories of related concepts to be inspected, reweighted, or ablated together.

To support reliable zero-shot scoring with vision-language models, we constrain the generated concepts to be visually grounded,

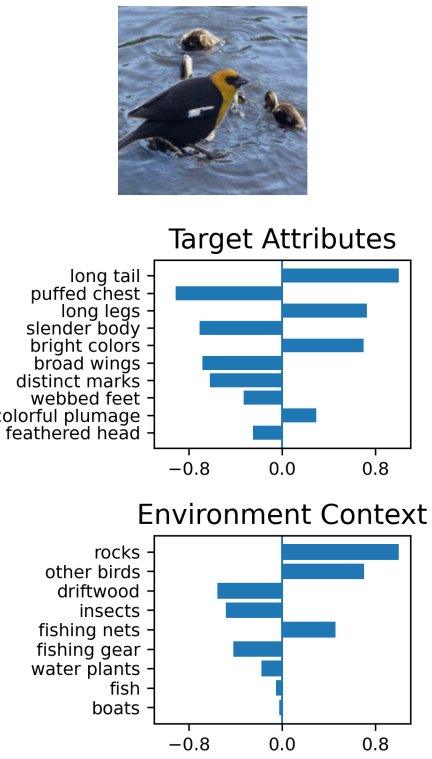


Figure 3: An example of concept-based attribution produced by the CRX model. For a single image (top), we visualize the most influential concepts within two semantic channels: Target Attributes and Environment Context (bottom). Each bar represents the signed contribution of a concept feature to the final logit. Positive contributions increase confidence in the predicted class, while negative contributions decrease it.

concrete, and identifiable from a single image. This helps ensure that concept scores behave as soft semantic features rather than ambiguous textual descriptors.

Importantly, not all channels are intended to be used symmetrically during prediction. Object-centric concepts (e.g., morphology or identity cues) can provide semantically meaningful evidence for classification. In contrast, environment context and co-occurring object concepts may reflect correlations that are not intrinsic to the object. We therefore include these contextual concepts explicitly to make potential spurious features visible in the representation. Downstream training can discourage reliance on these channels through masking (excluding channels entirely), sparsity (encouraging a small set of stable concepts), or coarse block-level gating that downweights an entire concept block when it is unreliable. In this way, the concept bank serves both as a source of interpretable evidence and as a structured "audit surface" for identifying and suppressing shortcut cues.

Zero-shot scoring with vision–language models. After constructing the concept bank, we estimate the presence of each concept in an image using a pretrained vision-language model (OpenCLIP RN50) [8]. Each concept phrase c is encoded into a normalized text embedding z_c, and each image x is encoded into a normalized image embedding z_x. Concept scores are then computed using scaled cosine similarity:

$$s_c(x) = \tau\, z_x^\top z_c, \tag{2}$$

where τ is CLIP's learned logit scale. Collecting these scores across all K concepts yields a concept vector $c(x) \in \mathbb{R}^K$ that provides a structured semantic description of the image.

This process is entirely zero-shot: the vision-language model is not trained or adapted to the target dataset. As a result, concept activations can be noisy or imperfect, particularly for fine-grained attributes or rare visual cues. Rather than interpreting these activations as ground-truth attributes, we treat them as soft semantic features that provide additional axes along which the classifier can form predictions. Even when individual concept detections are imperfect, they introduce semantically meaningful directions that can help disentangle object-centric evidence from contextual cues.

The concept bank metadata further allows selective control over which features influence training. Concepts can be masked entirely, filtered by channel, or downweighted indirectly through sparsity and block-level gating in the downstream classifier. This design enables the model to suppress unreliable or shortcut-related concepts while retaining useful semantic structure in the representation.

To illustrate the structure of the resulting concept representation, Figure 3 depicts concept activations for a representative image. Concepts categorized as "Target Attributes" produce scores reflecting the presence or lack of specific physical attributes, while scene-context concepts capture surrounding environmental cues. Because scores are contained via zero-shot vision-language similarity, they can be noisy and are interpreted as soft semantic features rather than ground-truth attributes. Even when imperfect, these activations introduce structured semantic directions that separate object evidence from contextual features.

Residual Feature Construction. Concept features cannot capture all causal or task-relevant variation in images. Many predictive cues are too fine-grained or abstract to be described by a finite set of human-readable concepts. We construct a residual feature representation from the ERM features to capture this additional information.

Let $c(x) \in \mathbb{R}^K$ denote the concept vector. We fit a ridge regression model on the validation split to predict ERM features from concepts:

$$W^* = \arg\min_W \|F - CW\|^2 + \lambda\|W\|^2, \tag{3}$$

where $F \in \mathbb{R}^{N\times d}$ stacks ERM features of N samples and $C \in \mathbb{R}^{N\times K}$ stacks concept features. For an example x, the concept-explainable component is $\hat{f}(x) = c(x)^\top W^* \in \mathbb{R}^d$. We define the residual

$$r(x) = f_\theta(x) - \hat{f}(x), \tag{4}$$

which captures components of the ERM representation that are not linearly explained by the concept space.

To control dimensionality and stabilize training, we apply principal component analysis (PCA) [1] to the residual vectors computed on the validation set. PCA is fit on the collection of residuals $r(x)$ across validation examples to identify directions of highest variance in the residual feature space. We retain the top k principal components and obtain a projection matrix P, whose columns form an orthonormal basis for the dominant residual directions. We define the reduced residual representation as:

$$r'(x) = P^\top r(x) \in \mathbb{R}^k. \tag{5}$$

This step compresses the high-dimensional residual vector while reducing noise and redundancy. Each residual principal component represents a recurring pattern in the ERM feature space not explained by the concept features.

Combined Representation. The final feature representation used for training is the concatenation of the frozen ERM features with concept and residual features:

$$h(x) = [\, f_\theta(x)\ ;\ c(x)\ ;\ r'(x)\,]. \tag{6}$$

The concatenation of multiple sources of information for improved representation learning has been explored in prior work, such as Chroma-VAE [38], which disentangles features in a similar manner. $h(x)$ preserves the full ERM feature space while adding interpretable, semantically structured dimensions (concepts), while preserving unstructured but potentially causal information not explained by the concept space (residuals). Together, these features expand the hypothesis space available to robust training without constraining the model to rely only on predefined attributes.

3.4 Expanded Feature Classification

Given the expanded representation $h(x) = [f_\theta(x); c(x); r'(x)]$, we train a lightweight classifier while keeping the ERM backbone frozen. Concretely, we optimize only the parameters of a multilayer perceptron classifier g_ϕ consisting of two hidden layers on top of the fixed features:

$$\hat{y} = g_\phi(h(x)). \tag{7}$$

We encourage sparse reliance on feature dimensions via an ℓ_1 penalty on the classifier weights:

$$\mathcal{L}(\phi) = \mathbb{E}\big[\ell_{\mathrm{CE}}(g_\phi(h(x)), y)\big] + \lambda_{\mathrm{clf}}\|\phi\|_1, \tag{8}$$

where ϕ denotes the weights of classifier g_ϕ associated with the final feature representation, and λ_{clf} represents the regularization hyperparameter for the classifier, controlling the sparsity of the

model's weights. Intuitively, because concept features are semantically aligned and grouped into channels, sparsity encourages the classifier to concentrate on a small subset of more stable features.

Computational Overhead. CRX adds limited trainable capacity beyond the ERM backbone. The ERM encoder and CLIP concept scorer are frozen, and Stage 3 trains only the classifier g_ϕ on the expanded representation. The main additional computation comes from one-time feature construction: caching CLIP concept scores, caching ERM features, fitting the ridge map from concepts to ERM features, and projecting residuals with PCA. Once these features are cached, training the final classifier is lightweight relative to training the ERM backbone. At inference time, CRX requires computing the ERM representation, concept scores, and the residual projection before applying g_ϕ, unless these features are precomputed.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on three spurious correlation benchmarks in vision: Waterbirds [28], CelebA [25], and CheXpert[16]. Waterbirds is a binary classification task where the target label is bird type, *landbird* versus *waterbird*, and the spurious attribute is background, *land* versus *water*. CelebA is a binary attribute classification task where the target label is *Blond Hair*, and the spurious attribute is gender. CheXpert is a chest X-ray classification task where the target label is *Finding* versus *No Finding*, and subgroups are defined by the intersection of race and gender metadata. Table 1 reports the group counts for each dataset, illustrating the subgroup imbalance present in these benchmarks. For all datasets, group labels are used only for evaluation of worst-group accuracy and are not used during training, validation, or model selection.

Backbone and Implementation Details. For all methods, we use a ResNet-50 backbone pretrained on ImageNet-1k as the ERM feature extractor. CRX freezes this ERM encoder and trains only the lightweight classifier on the expanded representation. Waterbirds and CelebA metadata are derived from the official dataset releases following the standard group definitions used in prior spurious correlation benchmarks. For CelebA, the target attribute is *Blond Hair*, and gender is used only to define evaluation groups. All processed metadata files and scripts for constructing the experimental splits are included with the released code to support reproducibility.

Baselines. We compare against representative approaches for robustness under spurious correlation, including ERM [32], GroupDRO [28], DFR [20], CRT [18], and JTT [24]. These methods cover reweighting, post-hoc adjustment, and group-robust optimization strategies that operate within a fixed learned representation.

Attribute Availability. Crucially, our method and all baselines are evaluated in the most challenging and realistic setting: group or environment labels are not available during training or validation. Models are trained without group supervision, and model selection does not use worst-group validation metrics. As a result, methods that would otherwise use group information instead function using class-based supervision. This setting prevents explicit group-aware optimization and forces methods to rely solely on learned representations.

Concept Bank Construction. For all experiments, CRX generates its concept vocabulary using only the task class names, such as "landbird" and "waterbird" for Waterbirds, rather than manually curated concept annotations or group labels. This represents a minimal-information setting in which the method has access to task semantics but not explicit spurious-attribute supervision. If richer task descriptions, metadata, or expert knowledge are available, they can be incorporated into the same concept-generation procedure without significant changes to the framework.

Table 1: Group counts for each dataset. Training splits exhibit subgroup imbalance, reflecting the spurious correlations that worst-group accuracy is designed to evaluate.

Group	Train	Val	Test
Waterbirds			
Landbird + Land	3498	467	2255
Landbird + Water	184	466	2255
Waterbird + Land	56	133	642
Waterbird + Water	1057	133	642
Total	4795	1199	5794
CelebA			
Not Blonde + Female	71629	8535	9767
Not Blonde + Male	66874	8276	7535
Blonde + Female	22880	2874	2480
Blonde + Male	1387	182	180
Total	162770	19867	19962
CheXpert			
Finding + White + Male	51606	6873	10420
Finding + White + Female	33676	4469	6772
Finding + Black + Male	4051	552	796
Finding + Black + Female	3877	513	783
Finding + Other + Male	33604	4484	6763
Finding + Other + Female	23407	3126	4696
No Finding + White + Male	5446	734	990
No Finding + White + Female	3490	487	661
No Finding + Black + Male	543	61	123
No Finding + Black + Female	506	71	94
No Finding + Other + Male	3912	518	740
No Finding + Other + Female	2975	392	581
Total	167093	22280	33419

Model Selection. For each algorithm, we perform a random search over 16 hyperparameter configurations. The best model is selected using validation average accuracy only without any group information. This protocol ensures that improvements in worst-group accuracy arise from better representations and inductive biases rather than oracle model selection.

Metrics. We report both average test accuracy (Mean) and worst-group accuracy (WGA). WGA measures the model's accuracy on the lowest-performing predefined group (defined by the label y and spurious attribute a). Formally,

$$\mathrm{WGA}(g_\phi) := \min_{(y,a)\in\mathcal{Y}\times\mathcal{A}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{(y,a)}}\left[\mathbb{1}\{g_\phi(x)=y\}\right], \quad (9)$$

where $\mathcal{D}_{(y,a)}$ denotes the test distribution restricted to the group with label y and spurious attribute a from the attribute set $\mathcal{A}$, and $\mathbb{1}\{g_\phi(x)=y\}$ is the 0-1 loss. A high worst-group accuracy indicate that the classifier is robust to spurious correlations and can fairly predict samples from different groups.

More details of the implementation are provided at: https://doi.org/10.5281/zenodo.20467988

Table 2: Test performance comparison across datasets without group labels. Best run per algorithm selected by validation mean accuracy over 16 hyperparameter trials.

	Waterbirds [28]		CelebA [25]		CheXpert [16]	
Algorithm	WGA (%)	Mean (%)	WGA (%)	Mean (%)	WGA (%)	Mean (%)
ERM	71.03	89.45	46.67	95.54	9.80	90.76
GroupDRO	70.72	89.61	43.33	95.70	55.66	87.34
CRT	64.02	91.11	65.56	94.82	34.04	87.03
DFR	83.33	93.79	44.44	95.53	0.00	90.46
JTT	69.47	89.94	32.22	95.93	16.06	90.88
CRX	**90.51**	91.15	**71.67**	93.47	**62.77**	84.43

Table 3: Ablation of representation expansion components without group labels.

	Waterbirds [28]		CelebA [25]	
Algorithm	WGA (%)	Mean (%)	WGA (%)	Mean (%)
ERM	71.03	89.45	46.67	95.54
+Concepts	88.91	90.16	70.00	94.00
+Residuals	89.10	90.78	71.11	93.67
+Both (CRX)	**90.51**	91.15	**71.67**	93.47

Table 4: Ablation comparing direct vision-language representations.

	Waterbirds [28]		CelebA [25]	
Algorithm	WGA (%)	Mean (%)	WGA (%)	Mean (%)
CLIP Direct	69.24	70.28	45.84	45.92
CLIP + Classifier	00.00	22.19	00.00	13.33
CRX	**90.51**	91.15	**71.67**	93.47

4.2 Main Results

As shown in Table 2, CRX substantially improves worst-group accuracy across all three datasets while maintaining competitive average performance. On each dataset, our method achieves the highest worst-group accuracy among all baselines. Compared to prior post-hoc approaches such as DFR, CRT, and JTT, which operate within a fixed ERM feature space, CRX consistently achieves stronger worst-group performance. On Waterbirds and CelebA, these gains are achieved with competitive average accuracy, while on CheXpert, CRX trades some average accuracy for a substantial improvement in worst-group accuracy.

These improvements can be understood in terms of how the expanded representation changes the features available for prediction. The ERM-trained backbone produces a high-dimensional feature space that entangles label-relevant and spurious information. Post-hoc methods that focus on reweighting or selecting among these existing directions can struggle when stable features are weak or difficult to separate from contextual features. CRX addresses this limitation by explicitly introducing additional, structured feature directions. Concepts provide semantically meaningful measurements of object attributes, scene elements, and contextual features, making key features more explicit and easier to isolate, while residual features ensure that useful predictive content is not lost. The gains in worst-group accuracy suggest that this expanded space promotes stable decision boundaries that perform well across groups, even without access to group labels.

Overall, the results indicate that CRX improves robustness across both standard spurious-correlation benchmarks and a larger clinical dataset. While the average-accuracy tradeoff varies by dataset, the consistent WGA gains suggest that representation expansion increases access to stable predictive features that are difficult to exploit using the ERM representation alone.

4.3 Ablations

We focus the ablation analysis on Waterbirds and CelebA, the two standard spurious-correlation benchmarks used for controlled component analysis. CheXpert is included in the main results as a larger clinical evaluation of the full CRX pipeline.

Concept and Residual Contribution. To isolate the contributions of different parts of the expanded representation, we compare models trained on (1) ERM features only, (2) ERM features augmented with concepts, (3) ERM features augmented with residuals, and (4) the full expanded representation (concepts + residuals).

This ablation directly tests whether robustness gains arise from representation expansion itself and whether concepts and residuals provide complementary information. Table 3 quantifies the contribution of each component. Augmenting ERM features with either concepts or residuals alone yields significant gains in worst-group accuracy, indicating that representation expansion itself is a primary driver of robustness gains. Notably, each component independently closes most of the gap between ERM and the full CRX model. This suggests that both concept and residual features introduce useful predictive structure that is not readily accessible within the original ERM feature space. The full CRX representation achieves the strongest worst-group accuracy overall, indicating that these two sources of information are complementary rather than redundant.

Robustness to Noisy Concept Scores. To evaluate whether CRX is sensitive to noisy concept activations, we perturb the test-split concept scores of a model trained under CRX with Gaussian noise, shown in Table 5. The added noise is scaled by each concept's

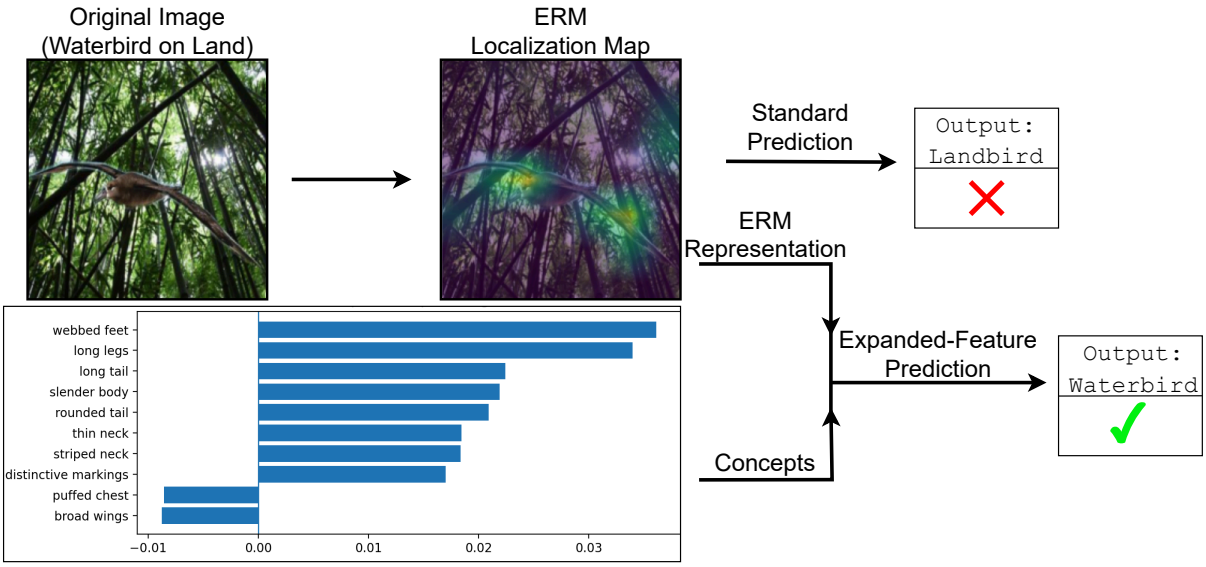


Figure 4: A case study depicting how concept features can influence prediction. The top pathway shows the standard ERM prediction based on the learned representation and its localization map obtained using Grad-CAM [29], where warmer colors indicate image regions that contribute more strongly to the model's decision. The bottom pathway shows prediction from the expanded representation, where the ERM representation is supported by concept features, providing additional concept scores from the "Target Attributes" channel and residual features that capture remaining semantic information. The bar length for each concept corresponds to its contribution score, computed as the product of the concept activation (from CLIP) and the classifier weight, indicating the signed contribution of that concept to the final class logit.

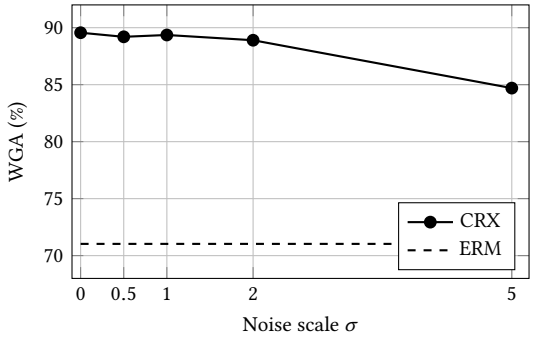


Figure 5: Ablation evaluating CRX robustness to noisy concept scores on Waterbirds [28]. Dashed lines indicate ERM baseline WGA values.

empirical standard deviation on Waterbirds. Worst-group accuracy remains stable under moderate noise, changing from 89.56% with clean concept scores to 89.36% at $\sigma = 1.0$ and 88.90% at $\sigma = 2.0$. Even under severe noise ($\sigma = 5.0$), CRX still retains 84.70% WGA.

Vision-language Baselines. We further assess whether the gains from CRX arise simply from using vision-language representations from CLIP models. Table 4 compares CRX against two baselines that rely directly on CLIP features: the first uses zero-shot CLIP predictions obtained via similarity between image embeddings and class text prompts. The second trains a simple classifier on frozen CLIP image embeddings using the same group-free training protocol as our method.

Zero-shot CLIP produces similar worst-group and average accuracies, indicating relatively balanced performance across groups. However, overall accuracy remains substantially lower than ERM and CRX. This behavior is consistent with the fact that zero-shot CLIP is not adapted to the training distribution and is therefore less sensitive to dataset-specific shortcuts, although it is also less aligned with the task. Training a classifier directly on frozen CLIP embeddings further results in collapse to predicting a single label, yielding zero worst-group accuracy. Together, these results suggest that CLIP features alone are insufficient for stable supervised learning in this setting.

4.4 Qualitative Analysis

Case Study. We examine a failure case to understand how representation expansion affects predictions. Figure 4 shows predictions on a Waterbirds dataset example in which a waterbird appears on land, a minority-group setting. Under ERM, the Grad-CAM [29] localization map highlights a few regions of the bird, suggesting the model attends to object-relevant structure. However, it also highlights a distant background region, indicating sensitivity to contextual cues. The model ultimately incorrectly predicts "Landbird," consistent with reliance on the surrounding scene for predictive cues.

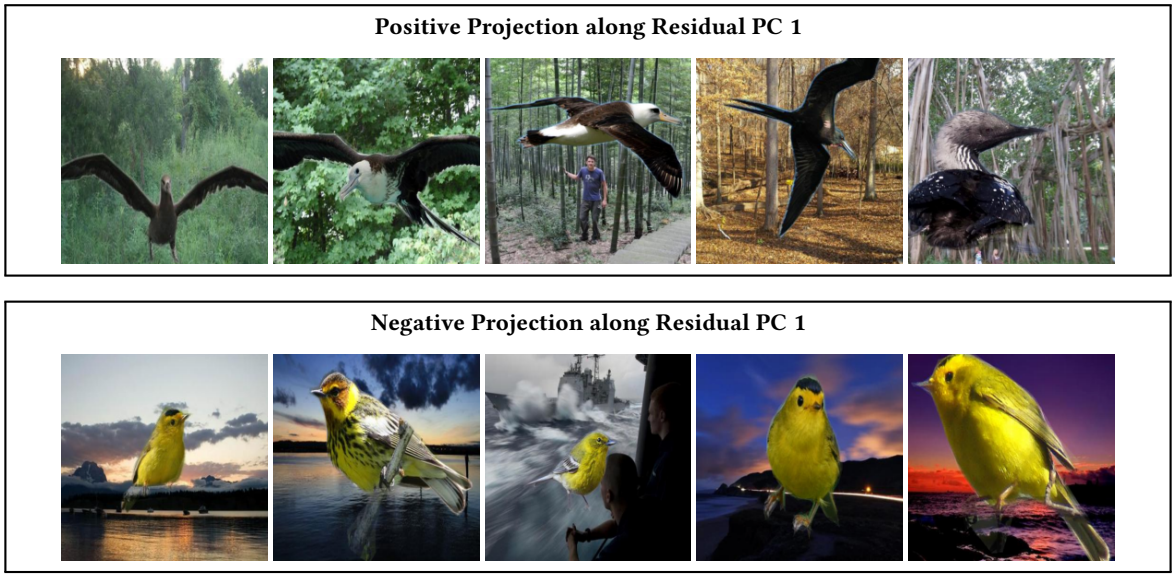


Figure 6: Examples with high activations along the first residual principal component (PC 1). The top row shows samples with large positive projections into this direction, while the bottom row shows samples with large negative projections, illustrating contrasting visual patterns captured by this residual space.

When concept features are introduced, the prediction changes to the correct label, "Waterbird." The concept features provide semantic evidence aligned with object identity. Morphology-related concepts such as "webbed feet" or "slender body" receive strong positive activations, supplying explicit features tied to the bird's physical characteristics rather than the environment. These concept-based cues help shift the decision away from background context and toward object-centered evidence. This expanded representation provides additional predictive structure that allows the classifier to rely less on contextual shortcuts and more on features that generalize across environments, resulting in the correct classification.

Residual Principal Components. To analyze what information is captured beyond the concept space, we inspect examples with strong positive and negative projections along the first principal component direction in the residual feature space (PC 1), shown in Figure 6. Images with strong positive projections along PC 1 contain large, monochrome birds, often in dynamic poses. In contrast, images with strong negative projections contain small, perched, bright yellow birds. This clear distinction supports that residual components capture meaningful, predictive visual structure, which can improve performance by providing additional information beyond the predefined concept space. However, because these directions are learned without semantic constraints, they may also encode spurious attributes, such as background trees in the positively projected examples, or clouds and water in the negatively projected ones. This highlights the complementary roles of the two components in the expanded representation: residual features provide additional predictive information that may be entangled with contextual cues, while concept features include structured, semantically grounded attributes that help anchor the decision.

5 Conclusion

We presented Concept-Residual eXpansion (CRX), a representation-centric framework for improving robustness to spurious correlations without group annotations. Rather than modifying the training objective alone, CRX augments an ERM-learned representation with structured concept features derived using CLIP, and also extracts complementary residual features that preserve visual information not captured by the concept space. A lightweight classifier trained on this expanded feature space will then provide robust model predictions given the instance-agnostic concept features, which can generalize reliably across groups. In experiments, CRX consistently improves worst-group accuracy while maintaining competitive average performance. Ablation results demonstrate that both concepts and residuals independently contribute to robustness, and that CRX remains stable under noisy concept activations. Additionally, comparisons with direct vision-language baselines indicate that CLIP-derived features are most effective when used to augment, rather than replace, strong ERM representations. Together, these findings highlight the potential of representation expansion as a promising and underexplored axis for improving robustness to spurious correlations.

Acknowledgment

This work is supported in part by the US National Science Foundation (NSF) and the National Institute of Health (NIH) under grants IIS-2106913, IIS-2538206, IIS-2529378, IIS-2500341, OAC-2530655, and R01LM014012-01. Any recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIH or NSF.

References

[1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.

[2] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. 2024. Zero-Shot Robustification of Zero-Shot Models. arXiv:2309.04344 [cs.LG] https://arxiv.org/abs/2309.04344

[3] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9690–9698.

[4] Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. 2024. Unsupervised Concept Discovery Mitigates Spurious Correlations. arXiv:2402.13368 [cs.LG] https://arxiv.org/abs/2402.13368

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

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

[7] Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. 2023. Why does throwing away data improve worst-group error?. In *International Conference on Machine Learning*. PMLR, 4144–4188.

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2818–2829.

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

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

[11] Yujin Han and Difan Zou. 2024. Improving group robustness on spurious correlation requires preciser group inference. *arXiv preprint arXiv:2404.13815* (2024).

[12] Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, and Jianhua Yao. 2022. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems* 35 (2022), 37704–37718.

[13] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[14] Haoyang Hong, Ioanna Papanikolaou, and Sonali Parbhoo. 2025. Do regularization methods for shortcut mitigation work as intended? arXiv:2503.17015 [cs.LG] https://arxiv.org/abs/2503.17015

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

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[17] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. 2022. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems* 35 (2022), 38516–38532.

[18] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217* (2019).

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2668–2677. https://proceedings.mlr.press/v80/kim18d.html

[20] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937* (2022).

[21] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

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

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

[24] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[26] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. 2022. Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation. arXiv:2204.02070 [cs.LG] https://arxiv.org/abs/2204.02070

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

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

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[30] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. 2021. Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems* 34 (2021), 16846–16859.

[31] Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. 2023. Group robust classification without any group information. *Advances in Neural Information Processing Systems* 36 (2023), 56553–56575.

[32] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* 10, 5 (1999), 988–999.

[33] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems* 34 (2021), 16196–16208.

[34] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *Advances in neural information processing systems* 34 (2021), 2215–2227.

[35] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3091–3100.

[36] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. 2023. Discover and Cure: Concept-aware Mitigation of Spurious Correlation. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 37765–37786. https://proceedings.mlr.press/v202/wu23w.html

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

[38] Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. 2022. Chroma-vae: Mitigating shortcut learning with generative classifiers. *Advances in Neural Information Processing Systems* 35 (2022), 20351–20365.

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

[40] W Ye, L Jiang, E Xie, G Zheng, Y Ma, X Cao, D Guo, D Qi, Z He, Y Tian, et al. 2024. The clever Hans mirage: A comprehensive survey on spurious correlations in machine learning. *arXiv: 2402.12715* (2024).

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

[42] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. 2021. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5372–5382.

[43] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. 2023. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16036–16047.