

Disentangling Spurious Correlations: How IRM Handles Diverse Shortcuts in Text Classification

Paul Grunenwald^{1,2*} Benoît Crabbé² Guillaume Wisniewski² Pirmin Lemberger¹

¹onepoint, Paris, France

²Laboratoire de Linguistique Formelle, Université Paris Cité, CNRS, Paris, France

Abstract

In NLP, models trained via Empirical Risk Minimization (ERM) often exploit spurious correlations, such as dialectal markers, at the expense of robust linguistic features, thereby degrading their out-of-distribution (OOD) generalization. Invariant Risk Minimization (IRM) aims to suppress such shortcuts by identifying predictive relationships that remain stable across environments, but its practical utility for NLP remains debated. We use a causality-driven framework to evaluate IRM under three shortcut-generating mechanisms: label-generated features, confounding, and selection bias. Our results show that IRM successfully rejects shortcuts across the evaluated causal and anti-causal structures, including under limited environmental variation and increasing label noise. We then extend our analysis beyond controlled synthetic artifacts to two less constrained settings: gender-profession associations in Bias in Bios and semantically loaded actor-name shortcuts in ActorCorr. These experiments show that invariant learning can mitigate more realistic textual shortcuts, while semantically meaningful features require careful calibration of the invariance penalty.

1 Introduction

Predictive success in NLP often masks a vulnerability: models frequently rely on dataset-specific biases rather than genuine robust linguistic features. Under the ERM paradigm (Vapnik, 2013), models exploit statistical shortcuts, such as negation words (Gururangan et al., 2018), syntactic heuristics (McCoy et al., 2019), or demographic dialectal markers (Blodgett et al., 2020), leading to catastrophic performance drops when encountering real-world distribution shifts where these biases no longer hold.

To mitigate this lack of robustness, several strategies have been proposed. Domain adaptation meth-

*Contact: paul.grunenwald@etu.u-paris.fr

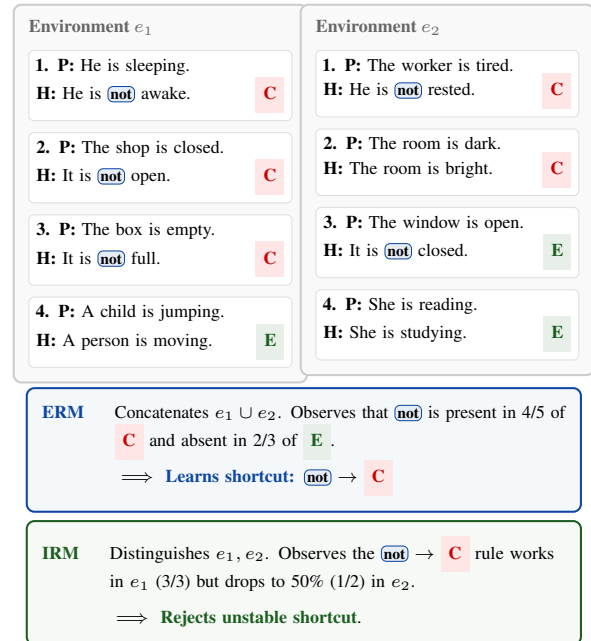


Figure 1: **Optimization dynamics under multi-environment training.** On a binary classification task, ERM exploits a global spurious correlation, while IRM leverages environment boundaries to detect its instability and discard it. (C: Contradiction, E: Entailment).

ods (Ben-David et al., 2010; Ganin et al., 2016) and sample reweighting techniques (Shimodaira, 2000) aim to align distributions or correct known selection biases. However, these techniques face a major limitation: they generally require prior knowledge of the test distribution, either through access to unlabeled target data or by explicitly identifying the specific confounding variables to be rebalanced. In many real-world NLP applications, this target distribution is unknown or subject to unpredictable shifts. In this context, IRM (Arjovsky et al., 2019) has emerged as a powerful theoretical alternative. The core originality of IRM lies in its ability to operate without any information regarding the target distribution. By leveraging data partitioned into multiple distinct training *environments*, IRM

enforces a strict learning constraint. It compels the representation to discard features whose predictive validity fluctuates, isolating only those stable, invariant correlations that yield simultaneously optimal classifiers across all environments. In doing so, IRM seeks to capture robust linguistic principles that guarantee optimal generalization, even to domains entirely unseen during training (Figure 1). Despite these theoretical promises, the practical effectiveness of IRM remains heavily debated. While large-scale benchmarks in machine learning have questioned its out-of-the-box superiority over ERM in standard domain generalization (Gulrajani and Lopez-Paz, 2021), evaluations on specific NLP tasks have also tempered expectations (Koh et al., 2021), notably regarding its capacity for robust bias mitigation (Dranker et al., 2021). To understand why cross-environment invariant methods lack effectiveness in NLP, we must move beyond treating spurious correlations as an undifferentiated problem. A non-causal correlation between a spurious textual variable and the label Y can originate from distinctly different mechanisms: a common cause, a selection bias, or a label-generated feature (Pearl, 2009). As these distinct generative processes result in structurally different distributional shifts, investigating how IRM responds to each specific spurious mechanism might help clarify the precise conditions required for its success in NLP. Furthermore, IRM’s core assumption, that spurious correlations must fluctuate across environments, raises questions about the required magnitude of variation and its behavior when the genuine task-relevant signal is degraded.

Our contributions. We provide a causal analysis of when IRM mitigates textual shortcuts. First, Section 4.2 distinguishes three shortcut-generating mechanisms, label-generated features, confounding, and selection bias, across causal and anti-causal tasks. Second, Sections 5.2 and 5.3 quantify how success depends on cross-environment variation and task-signal reliability. Third, Section 6 extends the analysis to naturally occurring gender-profession correlations and semantically loaded actor-name shortcuts, revealing an accuracy-invariance trade-off.

2 Background

2.1 Problem Setting: OOD Generalization

At the core of current challenges in NLP lies the issue of robustness to change, as models are rarely

deployed under linguistic conditions identical to those encountered during training. To formalize this heterogeneity, we adopt the concept of *environment* from the invariant learning literature (Peters et al., 2016; Arjovsky et al., 2019). Intuitively, an environment corresponds to a distinct data source, such as formal book reviews versus informal tweets, where the underlying task is identical, but the specific vocabulary, style, or local biases differ. Each environment e represents a specific generative context (such as a particular domain or demographic group) characterized by a joint probability distribution $P^e(X, Y)$ over texts X and labels Y . We denote by $\mathcal{E}_{\text{all}}$ the comprehensive set of all possible environments across which we evaluate a model’s ability to achieve OOD generalization. The goal is to find f that minimizes the risk $R^e(f)$ in the worst-case scenario across all possible environments:

$$\min_f \max_{e \in \mathcal{E}_{\text{all}}} R^e(f),$$

where $R^e(f) = \mathbb{E}_{(x,y) \sim P^e}[\ell(f(x), y)]$ represents the expected risk under environment e for a given loss function ℓ .

2.2 Empirical Risk Minimization and Its Limitations

The standard paradigm for predictive learning is ERM, which relies on a strict stationarity assumption: it presumes that the training data is independent and identically distributed (i.i.d.). It assumes that the global training distribution P_{train} , which is defined as the mixture of the joint distributions from all training environments $e \in \mathcal{E}_{\text{tr}}$, is strictly identical to the test distribution P_{test} (representing a target environment $e_{\text{test}} \notin \mathcal{E}_{\text{tr}}$). In practice, given a subset of training environments $\mathcal{E}_{\text{tr}} \subset \mathcal{E}_{\text{all}}$, ERM minimizes the average empirical risk $R_{\text{ERM}}(f)$ over a pooled dataset of size n :

$$\hat{f}_{\text{ERM}} = \arg \min_f \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) \quad (1)$$

By blindly minimizing this pooled loss, the model exploits all statistical correlations globally present in P_{train} . This is the core vulnerability of ERM: it fails to distinguish between robust features and spurious correlations. In an OOD scenario where $P_{\text{train}} \neq P_{\text{test}}$, these shortcuts vanish, inevitably causing the model’s performance to collapse.

2.3 Invariant Risk Minimization

To address this generalization failure, IRM leverages the distributional diversity across the available training environments $\mathcal{E}_{\text{tr}}$. Within this multi-environment framework, the fundamental intuition is to extract a unique feature representation Φ such that a single classifier w built on top of it remains simultaneously optimal across all contexts. The explicit role of Φ is to capture exclusively the stable correlations that persist across environments. By enforcing this strict invariance constraint, the model is compelled to discard environment-specific spurious correlations and rely solely on these robust features. Formally, IRM is framed as a bi-level optimization problem:

$$\begin{aligned} \min_{\Phi, w} \quad & \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(w \circ \Phi) \\ \text{s.t.} \quad & w \in \arg \min_{\tilde{w}} R^e(\tilde{w} \circ \Phi), \forall e \in \mathcal{E}_{\text{tr}}. \end{aligned} \quad (2)$$

Because this is an intractable bilevel optimization problem, a practical relaxation known as IRMv1 is typically used (Arjovsky et al., 2019; Choe et al., 2020; Drunker et al., 2021). In IRMv1, Φ is parameterized to directly output predictive logits, while the classifier w is reduced to a scalar dummy variable fixed at $w = 1$. Invariance is encouraged by penalizing the gradient of each environment-specific risk with respect to this dummy classifier:

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{\text{tr}}} \left[R^e(\Phi) + \lambda \|\nabla_{w|w=1} R^e(w \cdot \Phi)\|^2 \right]. \quad (3)$$

The dummy classifier w is not a trainable model component. Gradients of the penalty are propagated through Φ , whose parameters are optimized. The multiplier λ acts as a hyperparameter controlling the trade-off between standard empirical risk minimization and the invariance of the representation (enforced via the gradient penalty). We provide a sensitivity analysis of λ in Appendix A, where we examine its impact on model performance. Crucially, the success of IRM relies on the assumption that spurious correlations vary across $\mathcal{E}_{\text{tr}}$, while task-relevant predictive relationships remain invariant. This dependency motivates our environment construction strategy in Section 4, where we deliberately generate training sets by varying the strength of spurious correlations across environments, thereby creating the necessary heterogeneity to rigorously evaluate IRM’s invariance.

3 Related Work

Applications of invariant learning in NLP. Applications of invariant learning to NLP have produced mixed results. In the WILDS benchmark, Koh et al. (2021) evaluated IRM alongside several other methods for improving robustness to naturally occurring domain and subpopulation shifts, including NLP tasks, but found no consistent advantage over standard training. Focusing more directly on natural language inference, Drunker et al. (2021) similarly reported limited benefits when natural MNLI genres were treated as environments. However, IRM performed more favorably when environments were explicitly constructed to differ in the prevalence of a known bias, emphasizing the importance of environment definition. Other studies have adapted invariant learning more directly to NLP-specific objectives. Peyrard et al. (2022) introduced an invariant language-modeling objective and obtained clearer benefits in controlled bias-removal experiments than across broad, heterogeneous corpora. For demographic bias mitigation, Zhou et al. (2023) proposed Causal-Debias, which identifies bias-relevant factors as non-causal and label-relevant factors as causal, intervenes on the former across demographic groups, and applies an invariant learning loss during downstream fine-tuning. Their experiments show that this task-aware formulation can reduce stereotypical associations while preserving downstream performance. Taken together, these studies suggest that invariant learning is most effective when the environments or interventions are explicitly designed to expose the targeted bias. Our work complements these approaches by studying whether IRM’s behavior depends on the mechanism that generates the spurious correlation and by extending the evaluation from controlled synthetic shortcuts to naturally occurring and semantically loaded signals.

Alternative approaches to shortcut mitigation.

Other mitigation methods rely on different signals to identify shortcut-conflicting examples. Group DRO minimizes the worst-group risk but requires annotations for the relevant subpopulations (Sagawa et al., 2020). JTT removes this requirement by upweighting examples misclassified by an initial ERM model (Liu et al., 2021), whereas LfF uses a deliberately biased model and assumes that spurious features are easier to learn than task-relevant ones (Nam et al., 2020). Bias-model approaches similarly reweight examples according to

their compatibility with a known superficial heuristic (Utama et al., 2020). IRM uses a distinct source of supervision: rather than identifying biased examples directly, it suppresses predictive relationships that fail to support a common classifier across environments. It therefore does not require the shortcut to be explicitly annotated or easier to learn, but it does require the environment partition to expose variation in its association with the target.

4 Experiment Design

4.1 Model Architecture

All experiments share the same two-component architecture: a frozen DistilBERT encoder (Sanh et al., 2019) $\Phi_{\text{BERT}} : \mathcal{X} \rightarrow \mathbb{R}^{768}$, which extracts fixed sentence-level embeddings via masked mean pooling, followed by a single trainable linear head $W_\theta \in \mathbb{R}^{d_{\text{out}} \times 768}$, where $d_{\text{out}} = 1$ for binary classification and $d_{\text{out}} = C$ for C -class classification. To connect this architecture with the IRMv1 notation introduced in Section 2.3, we define the complete logit-producing map as $\Phi_\theta(x) = W_\theta \Phi_{\text{BERT}}(x)$. Thus, Φ_θ plays the role of the trainable classifier Φ in the IRMv1 objective, even though the pretrained encoder Φ_{BERT} itself remains frozen. Its trainable parameters θ correspond exclusively to the linear head W_θ . ERM and IRMv1 consequently optimize exactly the same model parameters W_θ and differ only in their training objective. ERM minimizes the standard classification loss over the pooled training examples, whereas IRMv1 retains the two training environments $\{e_1, e_2\}$ and optimizes the sum of the environment-specific risks augmented by the invariance penalty from Equation 3.

Keeping Φ_{BERT} frozen is a deliberate design choice. Because W_θ constitutes the only trainable component under both ERM and IRMv1, the two methods are compared within the same fixed pre-trained representation space and with an identical set of trainable parameters. This places our analysis in a controlled linear setting, in which any behavioral difference between ERM and IRMv1 can be attributed to the optimization objective rather than to encoder fine-tuning dynamics or differences in model capacity. This design therefore isolates the effect of the invariance objective, which is the primary object of study.

4.2 Environment Construction

Controlled shortcut injection makes it possible to manipulate a feature’s correlation with the label

Mechanism	Param.	e_1	e_2	OOD
Label-generated	p_c	0.70	0.90	0.00
Confounding	a	0.01	0.10	1.00
Selection bias	p_s	0.70	0.90	0.00

Table 1: Environment parameters shared across datasets. The OOD environment inverts the training-time shortcut. Parameter definitions and dataset-specific shortcut realizations are provided in Appendix F.

and directly measure the consequences of relying on it. Previous work has used this approach to study shortcut formation and evaluate model explanations (Bastings et al., 2022; Pezeshkpour et al., 2022). Following the environment-based formulation of IRM (Arjovsky et al., 2019), we construct two training environments in which a designated shortcut remains predictive but varies in strength, together with an OOD test environment in which its correlation with the label is inverted. Unlike prior protocols that treat injected artifacts as a single class of spurious correlations, we organize our environments according to the process generating the shortcut–label correlation. Following the causal terminology of Veitch et al. (2021), we distinguish label-generated features, confounding, and selection bias, using graphical representations inspired by their framework. For each mechanism, we vary only the structural equation or sampling component responsible for the shortcut while preserving the task-relevant relationship. Table 1 summarizes the shared environment parameters, while Figures 2, 3, and 4 illustrate the corresponding structures.

Label-Generated Feature (Figure 2). In this mechanism, $Y \rightarrow Z \rightarrow X_Y^\perp$: the target label Y generates a latent concept Z , which in turn produces the observable spurious feature $X_Y^\perp$. Mirroring the classic Colored MNIST setup Arjovsky et al. (2019), this label-dependent process can make the shortcut more predictive than the genuine task-relevant textual signal.

We implement this mechanism by inserting a class-associated token before neutral function words¹ throughout the text. The parameter $p_c^{(e)} = P^{(e)}(Z = Y \mid Y)$ controls the probability that the injected token agrees with the label. The OOD intervention reverses this correlation. To prevent the task-relevant signal from trivially dominating, we apply label noise at rate $\epsilon = 0.2$ before token injection.

¹For example, *the, a, an, of, in, to*.

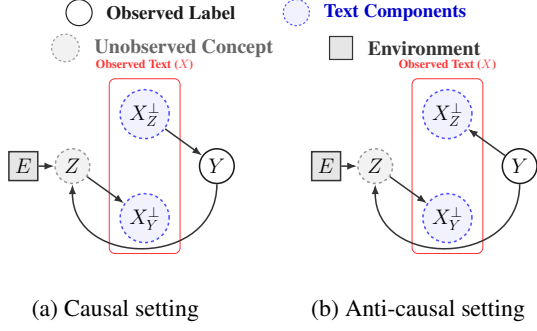


Figure 2: **Label-Generated Feature.** Spurious mechanism: $Y \rightarrow Z \rightarrow X_Y^\perp$. The red box indicates the observable text X .

Confounding (Figure 3). An unobserved confounder U influences both the target Y and a latent concept Z , opening the backdoor path $Y \leftarrow U \rightarrow Z \rightarrow X_Y^\perp$. This induces a non-causal statistical dependency that models can exploit as a predictive shortcut. The influence of U on Y remains fixed across environments, while only the fidelity of the proxy Z varies. This preserves the invariance conditions required by IRM, see Appendices B and C.

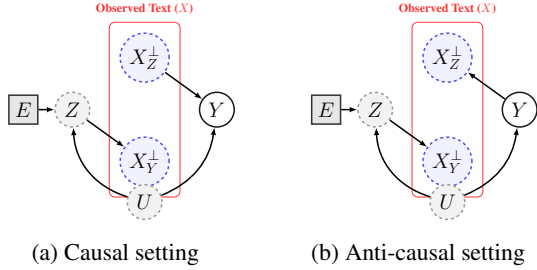


Figure 3: **Confounding.** Backdoor path via unobserved U .

Experimentally, we sample a binary confounder $U \sim \text{Ber}(0.5)$, which influences Y through a fixed, environment-independent mechanism. Its environment-dependent proxy is generated as $Z^{(e)} = U \oplus \text{Ber}(a^{(e)})$, where $\oplus$ denotes XOR. The parameter $a^{(e)}$ is the probability of flipping the value of U , so smaller values produce a more faithful shortcut. The OOD intervention completely reverses the proxy while leaving the relationship between U and Y unchanged. The resulting concept $Z^{(e)}$ determines the injected textual feature.

Selection Bias (Figure 4). An unobserved selection mechanism S is influenced by both Y and a latent feature Z , forming the collider $Y \rightarrow S \leftarrow Z$. The feature Z also generates the observable component $X_Y^\perp$. Training only on examples satisfying $S = 1$ induces a non-causal correlation between Y

and $X_Y^\perp$. Unlike the previous mechanisms, selection does not change their underlying values, but biases their observed co-occurrence frequencies.

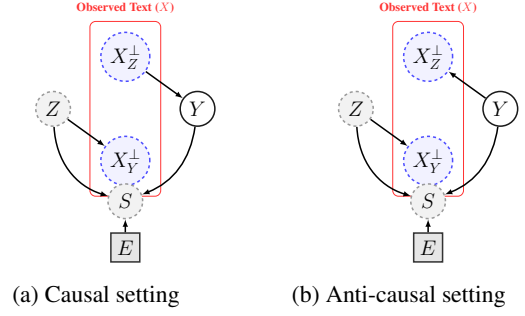


Figure 4: **Selection Bias.** Conditioned on selection collider S .

We simulate selection bias by subsampling each dataset according to the alignment between Y and a dataset-specific latent feature Z . Examples for which Z heuristically agrees with Y are called *aligned*, whereas the remaining examples are *misaligned*. The parameter $p_s^{(e)} = P^{(e)}(\text{aligned} \mid S = 1)$ controls the proportion of aligned examples retained in environment e . The OOD test set contains only misaligned examples, making the training-time correlation systematically misleading. The dataset-specific definitions of Z and alignment are provided in Appendix F.

4.3 Datasets

We evaluate the three shortcut-generating mechanisms on three balanced text-classification benchmarks covering causal and anti-causal tasks. Each corpus is partitioned into 80% training, 10% in-distribution validation, and 10% OOD test data. The training partition is divided equally between the two environments defined in Section 4.2.

Amazon Books Reviews (Hou et al., 2026)

(Causal, $X \rightarrow Y$): We predict whether a book review is unhelpful ($Y = 0$) or helpful ($Y = 1$) based on its text and user votes. The balanced corpus contains 100,000 reviews, including 80,000 training examples. We treat this task as causal because the review’s informativeness and quality influence its perceived helpfulness.

AG News (Zhang et al., 2015) (Causal, $X \rightarrow Y$):

We classify news articles into World, Sports, Business, or Sci/Tech ($Y \in \{0, 1, 2, 3\}$). The balanced corpus contains 127,600 articles, including 101,460 training examples. We treat this task as causal because the linguistic content of an article determines

its thematic category.

IMDB Genres (Maas et al., 2011) (Anti-causal, $Y \rightarrow X$): We classify movie plot descriptions as Thriller ($Y = 0$) or Romance ($Y = 1$). The balanced corpus contains 87,533 descriptions, including 70,027 training examples. We treat the task as anti-causal because the movie genre acts as a latent generative variable that influences the description’s vocabulary, style, and thematic content. Although IRM is often motivated through causal prediction, it can also identify invariant predictive relationships in anti-causal settings when class proportions remain stable across environments, as ensured by our balanced construction. Appendix D provides the formal argument and discusses the limitation under label shift.

5 Results

We evaluate the ability of IRM to identify robust predictors across nine experimental configurations (three spurious correlation mechanisms applied to three datasets, as summarized in Table 1). The aim is to quantify the robustness of invariance against the structural nature of the misleading correlation and environmental variability.

5.1 ID and OOD Performance

Tables 2 and 3 report, respectively, the in-distribution (ID) validation accuracy and the OOD test accuracy across all configurations.

As expected in Table 2, ERM achieves higher ID accuracy in all configurations, with IRM incurring losses of 2.2 to 11.7 percentage points. This reflects the cost of discarding unstable features that remain predictive within the training distribution.

Under shortcut inversion, IRM outperforms ERM in all nine dataset–mechanism configurations, with gains ranging from 6.9 to 58.1 percentage points. The improvement spans label-generated features, confounding, and selection bias, suggesting that shortcut rejection does not depend on a single generating mechanism. It also holds on the anti-causal IMDB task under the stable class proportions imposed by our balanced construction, consistent with the conditions discussed in Appendix D.

5.2 Sensitivity Analysis

To identify the minimal conditions for IRM to succeed, we varied the statistical disparity between training environments, defined as the gap in the strength of the spurious correlation across $\mathcal{E}_{\text{tr}}$ (Choe

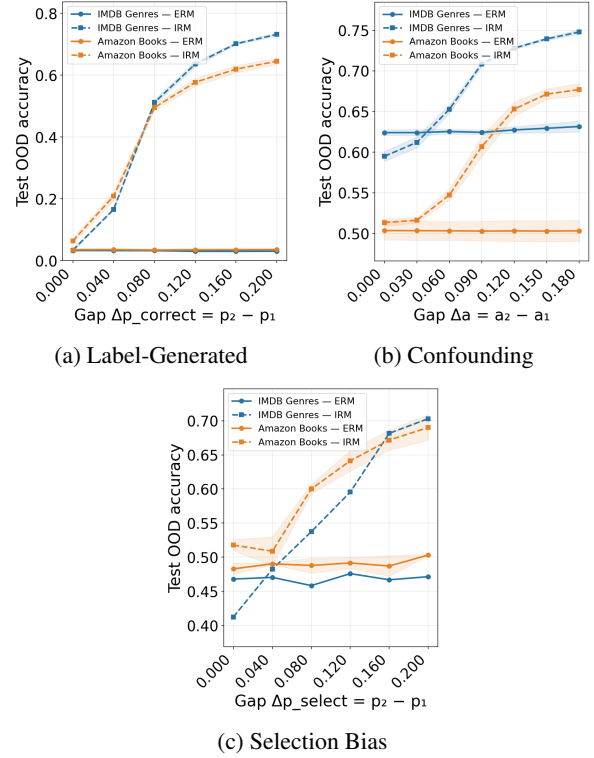


Figure 5: **Sensitivity analysis.** OOD accuracy as a function of the correlation gap for label-generated features (top-left), confounding (top-right), and selection bias (bottom-center).

et al., 2020). This disparity is the fundamental requirement for the invariance penalty: it creates the distributional instability that allows the model to distinguish unreliable shortcuts from stable causal correlation. Concretely, for a target gap Δ , environments are constructed symmetrically around a fixed center c : $p_1 = c - \Delta/2$ and $p_2 = c + \Delta/2$. For Label-Generated and Selection Bias, we fix $c = 0.85$ so that both environments maintain a globally strong spurious signal; only their differential varies (e.g., $\Delta = 0.20$ yields $p_1 = 0.75$, $p_2 = 0.95$). For Confounding, the center is $c = 0.05$ on the proxy-noise parameter a , with $\Delta \in [0, 0.18]$. At $\Delta = 0$, the two environments are statistically identical and IRM cannot detect any correlation instability, degenerating to ERM.

Our results (Figure 5) reveal that IRM can leverage even the slightest statistical disparity. While a minimal gap is often sufficient for the invariance penalty to begin discarding the misleading correlation in favor of a stable task-relevant signal, we observe that IRM’s OOD performance tends to increase as the correlation gap between environments widens. This sensitivity is an encouraging sign for real-world applications where domain differences

Spurious Mechanism	Amazon (Causal)	AG News (Causal)	IMDB (Anti-causal)
Label-Generated	$0.814 \pm 0.001 / 0.758 \pm 0.008$ (-0.056)	$0.810 \pm 0.000 / 0.769 \pm 0.004$ (-0.041)	$0.812 \pm 0.002 / 0.790 \pm 0.007$ (-0.022)
Confounding	$0.869 \pm 0.005 / 0.833 \pm 0.014$ (-0.036)	$0.862 \pm 0.003 / 0.745 \pm 0.002$ (-0.117)	$0.740 \pm 0.002 / 0.718 \pm 0.002$ (-0.022)
Selection Bias	$0.843 \pm 0.003 / 0.788 \pm 0.009$ (-0.055)	$0.926 \pm 0.001 / 0.881 \pm 0.004$ (-0.045)	$0.794 \pm 0.003 / 0.716 \pm 0.005$ (-0.078)

Table 2: In-distribution (ID) validation accuracy for ERM and IRM.

Spurious Mechanism	Amazon (Causal)	AG News (Causal)	IMDB (Anti-causal)
Label-Generated	$0.117 \pm 0.003 / 0.672 \pm 0.007$ (+0.555)	$0.417 \pm 0.011 / 0.798 \pm 0.007$ (+0.381)	$0.147 \pm 0.009 / 0.728 \pm 0.008$ (+0.581)
Confounding	$0.501 \pm 0.006 / 0.638 \pm 0.032$ (+0.137)	$0.244 \pm 0.010 / 0.699 \pm 0.014$ (+0.455)	$0.622 \pm 0.005 / 0.691 \pm 0.012$ (+0.069)
Selection Bias	$0.544 \pm 0.005 / 0.721 \pm 0.004$ (+0.177)	$0.756 \pm 0.008 / 0.836 \pm 0.010$ (+0.080)	$0.209 \pm 0.006 / 0.701 \pm 0.025$ (+0.492)

Table 3: Out-of-distribution (OOD) test accuracy for ERM and IRM.

may be subtle.

5.3 Robustness to Label Noise

We investigate the models’ behavior as the reliability of the task-relevant signal is progressively degraded. By gradually increasing the label noise rate ϵ , we artificially weaken the predictive power of the causal features relative to the spurious ones. This experiment is designed to determine whether a model continues to favor invariant features as the task-relevant signal becomes less reliable.

As shown in Figure 6, increasing ϵ progressively weakens the causal signal while leaving the shortcut configuration unchanged, making the spurious signal relatively more attractive for minimizing training loss. ERM succumbs to this bias. As the genuine signal fades, the model shifts its reliance to the shortcut, leading to catastrophic performance degradation at test time when the correlation is inverted. Conversely, IRM remains largely insensitive to label noise. IRM does not select features based on their relative predictive convenience, but strictly based on their structural stability. By rejecting correlations that vary across environments, IRM maintains its focus on the causal features, preferring a noisy yet invariant signal over a cleaner but unstable shortcut, even when the latter offers lower training loss.

6 Beyond Synthetic Shortcuts

The controlled experiments in Section 4 isolate the origin and strength of each spurious correlation.

We now consider two less controlled settings. Bias in Bios (De-Arteaga et al., 2019) contains a naturally occurring association between gender and profession, whereas ActorCorr (Eshuijs et al., 2025) introduces real entity names whose representations already carry semantic information from pretraining. These experiments test whether the behavior observed with neutral synthetic artifacts extends to more realistic shortcuts.

6.1 Naturally Occurring Gender Correlations

We use Bias in Bios, a corpus of 397,000 biographies annotated with profession and binary gender. We construct binary tasks pairing a male-typical profession with a female-typical profession and resample two training environments such that $P(\text{gender} = \text{canonical}(y) \mid Y = y)$ equals 0.99 and 0.80, respectively. This naturally occurring correlation is fully inverted at test time. We add a label-noise rate of $\epsilon = 0.25$ and train the same frozen DistilBERT classifier as in the controlled experiments. To examine whether semantic separability affects shortcut reliance, we compare the five semantically closest and the five most semantically distant eligible profession pairs. Each training environment contains at least 1,000 examples, and results are averaged over three seeds. Further details on environment construction, pair selection, and semantic-distance estimation are provided in Appendix G.

Across the ten profession pairs, IRM improves mean OOD accuracy from 43.35% to 78.78%, cor-

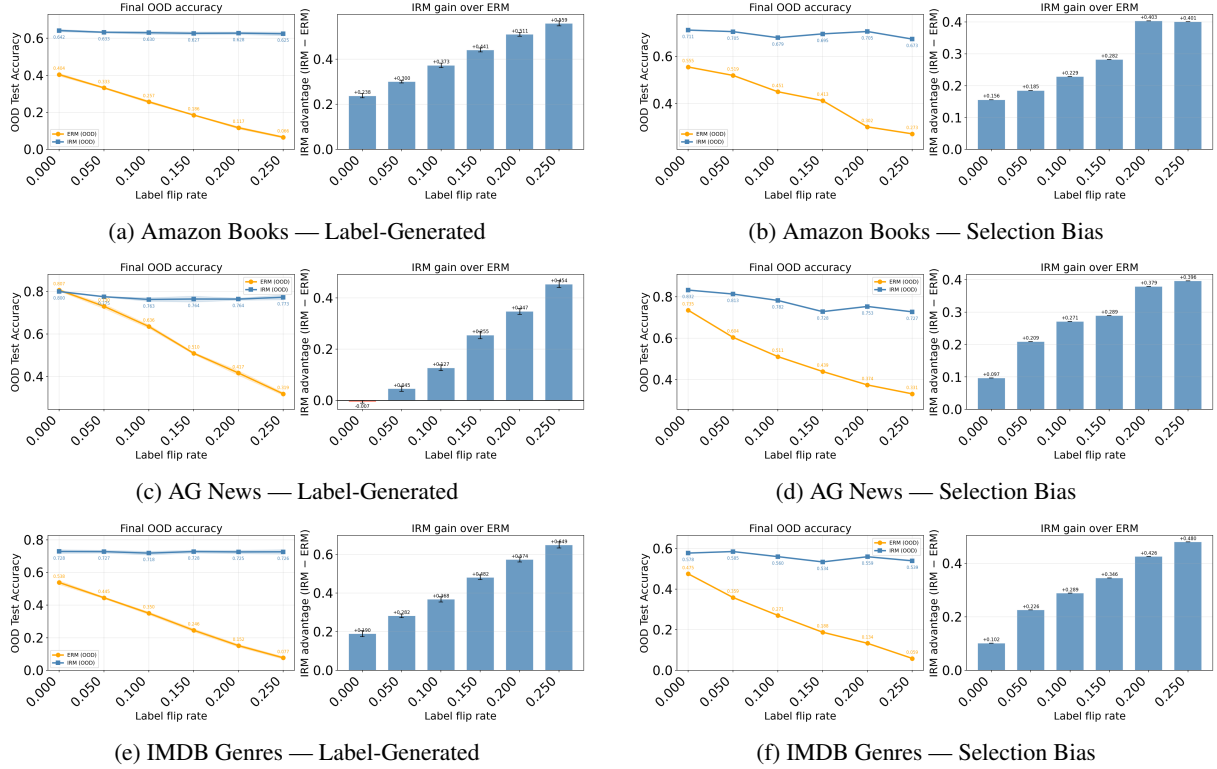


Figure 6: **Noise-sweep analysis.** OOD accuracy as a function of label noise. Rows correspond to Amazon Books (top), AG News (middle), and IMDB Genres (bottom). Columns correspond to label-generated features (left) and selection bias (right).

Profession pair	ERM	IRM	Δ
<i>Semantically close pairs</i>			
attorney/paralegal	8.4 ± 0.6	47.1 ± 0.5	$+38.8 \pm 0.7$
professor/psychologist	22.5 ± 0.5	79.6 ± 1.4	$+57.1 \pm 1.3$
chiropractor/psychologist	33.9 ± 3.0	77.4 ± 1.4	$+43.5 \pm 4.4$
chiropractor/nurse	41.9 ± 2.0	73.6 ± 3.2	$+31.7 \pm 5.2$
physician/nurse	35.0 ± 0.5	83.6 ± 0.9	$+48.6 \pm 1.1$
<i>Semantically distant pairs</i>			
dj/nurse	47.1 ± 2.4	91.7 ± 1.4	$+44.6 \pm 1.1$
rapper/psychologist	39.8 ± 2.3	79.4 ± 4.5	$+39.6 \pm 3.7$
dj/paralegal	74.9 ± 1.2	81.5 ± 14.1	$+6.6 \pm 13.2$
rapper/paralegal	77.3 ± 3.0	86.9 ± 3.5	$+9.6 \pm 0.5$
rapper/nurse	52.7 ± 1.8	87.0 ± 2.8	$+34.3 \pm 4.3$

Table 4: Bias in Bios OOD accuracy for semantically close and distant profession pairs. Results are mean $\pm$ standard deviation over three seeds; values are percentages.

responding to an average gain of 35.44 percentage points. The improvement is positive for every evaluated pair. The results are consistent with greater shortcut reliance when professions are semantically similar. For the five semantically closest pairs, ERM and IRM reach average OOD accuracies of 28.3% and 72.3%, respectively, yielding

a 43.9 percentage-point gain. For the five most semantically distant pairs, they reach 58.4% and 85.3%, yielding a smaller 26.9 percentage-point gain. As expected, when profession pairs are semantically close, profession-specific cues are less discriminative, making gender a comparatively easier and more attractive signal for separating the classes. For semantically distant pairs, other textual features already provide strong predictive cues, reducing the relative usefulness of gender. IRM’s average advantage is therefore 17.0 percentage points larger for semantically close pairs, primarily because ERM is less robust when profession-specific representations are harder to separate.

6.2 Semantically Loaded Entity Shortcuts

We next evaluate ActorCorr, which introduces actor names into IMDB reviews as semantically loaded sentiment shortcuts. At a fixed injection frequency of 5%, Morgan Freeman is associated with positive sentiment and Adam Sandler with negative sentiment. We construct two training environments with shortcut purities $\pi^{(e)} \in \{1.0, 0.85\}$, where $\pi^{(e)}$ is the probability that the injected actor agrees with the sentiment label. Unlike the neutral tokens

Model	λ	OOD $\uparrow$	ACAC $\downarrow$
ERM	–	0.739 ± 0.004	0.186 ± 0.002
IRM	10	0.745 ± 0.007	0.146 ± 0.002
IRM	100	0.704 ± 0.008	0.049 ± 0.002

Table 5: ActorCorr aggregate results over three seeds. IRM with $\lambda = 10$ achieves the highest pooled OOD accuracy, whereas IRM with $\lambda = 100$ achieves the lowest signed ACAC.

used in our controlled experiments, these names may carry associations inherited from pretraining. Each test review is evaluated under three conditions: *clean*, , corresponding to the original unmodified review; *correlated*, with an actor agreeing with the label; and *anti-correlated*, with an actor associated with the opposite label.

We report the Anti-Correlated Accuracy Change (ACAC; Eshuijs et al. (2025)), defined as the signed average of the class-specific accuracy difference between the original (clean) and anti-correlated conditions:

$$\text{ACAC} = \frac{1}{2} \sum_{y \in \{\text{pos}, \text{neg}\}} \left(\text{Acc}_{\text{clean}}^{(y)} - \text{Acc}_{\text{anti}}^{(y)} \right). \quad (4)$$

Higher positive values indicate greater sensitivity to misleading actor cues. Because ACAC is signed, class-specific effects may partially cancel (we analyze this in Appendix H).

Table 5 shows a clear trade-off between OOD performance and shortcut invariance. ERM achieves an OOD accuracy of 0.739 ± 0.004 with an ACAC of 0.186 ± 0.002 . With $\lambda = 10$, IRMv1 slightly improves OOD accuracy to 0.745 ± 0.007 while reducing ACAC to 0.146 ± 0.002 . Increasing the penalty to $\lambda = 100$ further lowers ACAC to 0.049 ± 0.002 , but OOD accuracy drops to 0.704 ± 0.008 . Thus, larger penalties reduce sensitivity to the spurious actor cue but can also hurt predictive performance, requiring a trade-off between robustness and OOD accuracy. The low ACAC at $\lambda = 100$ should nevertheless be interpreted cautiously because of partial class-specific cancellation (Appendix H).

Table 6 highlights the contrasting behavior of ERM and IRMv1 ($\lambda = 100$) across evaluation conditions. ERM relies heavily on the actor cue: its accuracy rises to 0.870 when the cue is correlated with the label but falls to 0.574 when it becomes misleading. In contrast, IRMv1 is less sensitive to this shift, with accuracies of 0.726 and 0.671 in the correlated and anti-correlated conditions, respec-

Metric	ERM	IRM
ID accuracy	0.755 ± 0.002	0.733 ± 0.003
Pooled OOD accuracy	0.739 ± 0.004	0.704 ± 0.008
Clean accuracy	0.759	0.716
Correlated accuracy	0.870	0.726
Anti-correlated accuracy	0.574	0.671
Spread $\downarrow$	0.296	0.055
ACAC $\downarrow$	0.186 ± 0.002	0.049 ± 0.002

Table 6: ActorCorr condition-level results over three seeds for ERM and IRMv1 with $\lambda = 100$. Spread is the difference between the highest and lowest condition-level accuracies.

tively, reducing the condition spread from 0.296 to 0.055. This robustness comes at the cost of lower clean and correlated accuracy, illustrating how invariant learning can suppress a semantically meaningful shortcut while sacrificing some predictive performance. The full penalty sweep and class-specific results are reported in Appendix H.

7 Conclusion

We systematically evaluated IRM across label-generated features, confounding, and selection bias. Under controlled shortcut shifts, IRM consistently improves OOD robustness across causal tasks and succeeds in anti-causal settings with stable class prevalence. As the task-relevant signal weakens, IRM also remains more robust than ERM by favoring invariant predictors. Beyond controlled shortcuts, IRM improves OOD accuracy across all Bias in Bios profession pairs, with larger average gains for semantically similar professions, although semantic distance does not fully explain shortcut reliance. On ActorCorr, IRM reduces sensitivity to semantically loaded actor cues, but reveals an accuracy–invariance trade-off. Overall, IRM’s effectiveness depends less on the shortcut’s structural origin than on whether training environments expose its instability while preserving a stable task-relevant mechanism. These findings also show the importance of targeted interventions and class-specific metrics, as pooled OOD accuracy may conceal shortcut reliance. Extending this analysis to naturally occurring environments and fully fine-tuned language models remains an important direction for future work.

Limitations

A first limitation concerns the controlled nature of our spurious mechanisms. Selection bias differs fundamentally from the other mechanisms considered here: it arises at the observation stage through conditioning on a collider, whereas confounding and label-generated features can be part of the underlying data-generating process itself. To isolate these mechanisms while retaining real-world text semantics, we introduce the corresponding shortcuts post hoc through controlled token injection. This design enables precise manipulation of the relevant causal relationships, but does not capture spurious correlations that emerge organically in natural text, such as environment-dependent shifts in style, lexical choice, or discourse. Our injected cues should therefore be viewed as controlled proxies for more diffuse real-world shortcuts, whose causal sources are substantially harder to isolate. A second limitation concerns the scope of IRM theory. Although IRM has been widely studied in classification settings (Arjovsky et al., 2019; Ahuja et al., 2020; Choe et al., 2020; Kamath et al., 2021), its original theoretical guarantees were developed primarily for regression. Extending invariance principles to classification remains an active research problem: subsequent work provides guarantees under more specific assumptions (Ahuja et al., 2021), while their applicability to modern Transformer-based models remains less well understood (Peyrard et al., 2022). A third limitation is that we freeze the pretrained encoder and train only a linear classification head. This choice deliberately places ERM and IRMv1 in the same fixed representation space, allowing us to isolate the effect of the optimization objective, but it does not reflect the fully fine-tuned setting commonly used with language models. This is particularly relevant because IRM can behave differently in highly overparameterized models, where additional capacity may permit solutions that satisfy the invariance penalty without yielding the intended OOD generalization benefits (Rosenfeld et al., 2021; Gulrajani and Lopez-Paz, 2021). Extending the present analysis to fully fine-tuned Transformers therefore remains an important direction for future work. Finally, our study is designed to compare causal shortcut mechanisms under controlled interventions rather than to benchmark IRM against the broader bias-mitigation literature. Evaluating the same mechanisms in less controlled real-world settings would

enable direct comparisons with representative alternatives such as Group DRO (Sagawa et al., 2020), JTT (Liu et al., 2021), Learning from Failure (LfF) (Nam et al., 2020), and bias-model-based reweighting (Utama et al., 2020). Such comparisons would help clarify when the explicit environment supervision used by IRM provides advantages over methods based on group annotations or on identifying biased examples from model behavior.

Ethical Considerations

We exclusively utilize publicly available benchmarks in strict accordance with their original licenses for research purposes. Our methodology involves the artificial synthesis of environments through token injection; this process does not involve the collection of new personally identifiable information or the use of private user data. To minimize the environmental footprint of our extensive sensitivity analyses, we conducted all experiments using DistilBERT (Sanh et al., 2019), a smaller and more energy-efficient architecture than standard large-scale models.

References

- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.
- Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450.
- Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. 2020. Invariant risk minimization games. *International Conference on Machine Learning*, pages 145–155.
- Martin Arjovsky. 2020. *Out of Distribution Generalization in Machine Learning*. Ph.D. thesis, New York University.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.
- Jean-Thomas Baillargeon and Luc Lamontagne. 2022. Assessing the impact of sequence length learning on classification tasks for transformer encoder models. *arXiv preprint arXiv:2212.08399*.
- Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. “will

- you find these shortcuts?” a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1):151–175.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.
- Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. 2020. An empirical study of invariant risk minimization. *arXiv preprint arXiv:2004.05007*.
- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 120–128.
- Yana Drinker, He He, and Yonatan Belinkov. 2021. Irm — when it works and when it doesn’t: A test case of natural language inference. *Advances in Neural Information Processing Systems*, 34:18212–18224.
- Leon Eshuijs, Shihan Wang, and Antske Fokkens. 2025. Short-circuiting shortcuts: Mechanistic investigation of shortcuts in text classification. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 105–125.
- Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, and 1 others. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.
- Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. *International Conference on Learning Representations*.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.
- Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3251–3265.
- Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anti-causal learning for nlp. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9499–9513.
- Prithish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. 2021. Does invariant risk minimization capture invariance? *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, and 1 others. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.
- 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*, pages 6781–6792. PMLR.
- Andrew Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.
- Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.
- Judea Pearl. 2009. *Causality*. Cambridge University Press.

- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.
- Maxime Peyrard, Sarvjeet Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, Saurabh Tiwary, and Robert West. 2022. Invariant language modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5728–5743.
- Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron C Wallace. 2022. Combining feature and instance attribution to detect artifacts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1934–1946.
- Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2021. The risks of invariant risk minimization. *International Conference on Learning Representations*.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *NeurIPS EMC² Workshop*.
- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. In *International conference on machine learning*, pages 459–466.
- Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.
- Vladimir Vapnik. 2013. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.
- Dušan Variš and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257.
- Victor Veitch, Alexander D’Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28.
- Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241.

A Model Architecture and Experimental Setup

Before detailing our environmental synthesis framework, we formally define the predictive architecture shared by ERM and IRMv1 and clarify its correspondence with the notation introduced in Section 2.3. Let $\mathcal{X}$ denote the input text space and $\mathcal{Y}$ the downstream label space.

Encoder Architecture and Text Representation.

All experiments use a frozen DistilBERT encoder $\Phi_{\text{BERT}} : \mathcal{X} \rightarrow \mathbb{R}^{768}$. Specifically, we use DistilBERT (distilbert-base-uncased) (Sanh et al., 2019), a knowledge-distilled variant of BERT-base with 6 Transformer layers, 12 attention heads, and approximately 66 million parameters. DistilBERT retains 97% of BERT’s language understanding capabilities while being 40% smaller and 60% faster, making it well suited to the extensive sensitivity analyses conducted in this study. Texts are tokenized using the DistilBERT WordPiece tokenizer with a maximum sequence length of 256 tokens for AG News (Zhang et al., 2015), and 512 tokens for IMDB Genres and Amazon Books. For any input sequence $x \in \mathcal{X}$, the encoder produces a fixed sentence-level representation $z(x) = \Phi_{\text{BERT}}(x) \in \mathbb{R}^{768}$, obtained by masked mean pooling over the final hidden states of all non-padding tokens. The parameters of Φ_{BERT} remain frozen throughout all experiments.

Classification Head and IRMv1 Parameterization.

A single trainable linear head $W_{\theta} \in \mathbb{R}^{d_{\text{out}} \times 768}$ is placed on top of the frozen representation, with $d_{\text{out}} = 1$ for the binary Amazon Books and IMDB Genres tasks and $d_{\text{out}} = 4$ for AG News. The resulting predictive logits are

$\Phi_\theta(x) = W_\theta \Phi_{\text{BERT}}(x)$. We use Φ_θ deliberately to match the notation of IRMv1. In Equation 3, Φ denotes the parameterized mapping that directly outputs predictive logits. In our architecture, this mapping corresponds to the composition $W_\theta \circ \Phi_{\text{BERT}}$. Although Φ_{BERT} is fixed, Φ_θ remains trainable through the parameters θ of W_θ . Importantly, the matrix W_θ should not be confused with the scalar classifier w appearing in the IRMv1 objective. The former is the actual trainable classification head, whereas the latter is the dummy scalar variable at which the invariance penalty is evaluated:

$$\|\nabla_w|_{w=1} R^e(w \cdot \Phi_\theta)\|^2. \quad (5)$$

The dummy classifier w is not an independently optimized model component; gradients of this penalty are propagated through Φ_θ and therefore update W_θ . For the binary Amazon Books and IMDB Genres tasks, $\Phi_\theta(x) \in \mathbb{R}$ and we use binary cross-entropy (BCE) with logits, with predictions obtained by thresholding the logit at zero. For AG News, $\Phi_\theta(x) \in \mathbb{R}^4$ and we use standard multiclass cross-entropy loss.

Optimization and Hardware. Both ERM and IRM heads are trained using the Adam optimizer (Kingma and Ba, 2015) with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay of 10^{-4} . We use a batch size of 256. Each model is trained for 25,000 gradient steps. The learning rate and IRM penalty weight λ are kept identical for both ERM and IRM within each experimental configuration to rigorously isolate the effect of the objective function. All reported results are averaged over three random seeds (0, 1, 2). Finally, given the compact size of our frozen architecture, all experiments were conducted efficiently using a single NVIDIA RTX A5000 GPU with 24 GB of VRAM.

Hyperparameter Selection: IRM Penalty Weight λ . The penalty coefficient λ in IRMv1 (Equation 3) was selected via a grid search over $\lambda \in \{50, 75, 100, 125\}$, covering all dataset-mechanism combinations (3 datasets $\times$ 3 mechanisms $\times$ 4 noise levels $\times$ 2 correlation strengths, yielding 48 cells), with all other conditions held fixed (same seed, 25,000 gradient steps, learning rate 10^{-4}). Figure 7 shows the corresponding accuracy curves as a function of label noise, while Table 7 reports the underlying OOD accuracies. All IRM variants consistently outperform ERM, and

performance is broadly robust to the exact choice of λ ; $\lambda = 100$ achieves the best or near-best OOD accuracy in 31 out of 48 cells and is therefore fixed as the default for all main-paper experiments.

B Structural Equation Models and the Nature of Environments

In IRM theory, the algorithm exploits natural datasets collected from distinct temporal or spatial contexts to identify invariant correlations. However, since we synthesize our own environments by injecting misleading correlations, it is imperative to demonstrate the scientific validity of this procedure. This justification relies on the formal framework of Structural Equation Models (SEM).

Definition B.1 (SEM). An SEM $\mathcal{C} := (\mathcal{S}, N)$ governing a random vector $X = (X_1, \dots, X_d)$ is a set of structural equations:

$$S_i : X_i \leftarrow f_i(\text{Pa}(X_i), N_i)$$

where $\text{Pa}(X_i) \subseteq \{X_1, \dots, X_d\} \setminus \{X_i\}$ denotes the set of causal parents of X_i , and N_i are independent noise variables. We say that X_i causes X_j if $X_i \in \text{Pa}(X_j)$. The causal graph of X is obtained by drawing an edge from X_i to X_j if $X_i \in \text{Pa}(X_j)$, assuming the graph is a Directed Acyclic Graph (DAG).

Within this framework, an environment e is defined as the result of an intervention on the SEM (Peters et al., 2016; Arjovsky et al., 2019), consisting of replacing one or more of its structural equations to obtain a new SEM $\mathcal{C}^e = (\mathcal{S}^e, N^e)$. For IRM’s guarantees to be theoretically legitimate, these interventions must be “valid”, satisfying three criteria:

- The causal graph remains acyclic.
- The conditional expectation remains invariant: $\mathbb{E}[Y^e \mid \text{Pa}(Y)] = \mathbb{E}[Y \mid \text{Pa}(Y)]$.
- The variance $\mathbb{V}[Y^e \mid \text{Pa}(Y)]$ remains within a finite range.

Our protocol strictly respects these conditions. By injecting spurious variables without altering the intrinsic semantic link between the text and its label, we generate our training environments $\mathcal{E}_{\text{tr}}$ via valid interventions by mathematically modulating the strength of these spurious correlations. This approach ensures that any observed failure of IRM stems from a structural limitation of the

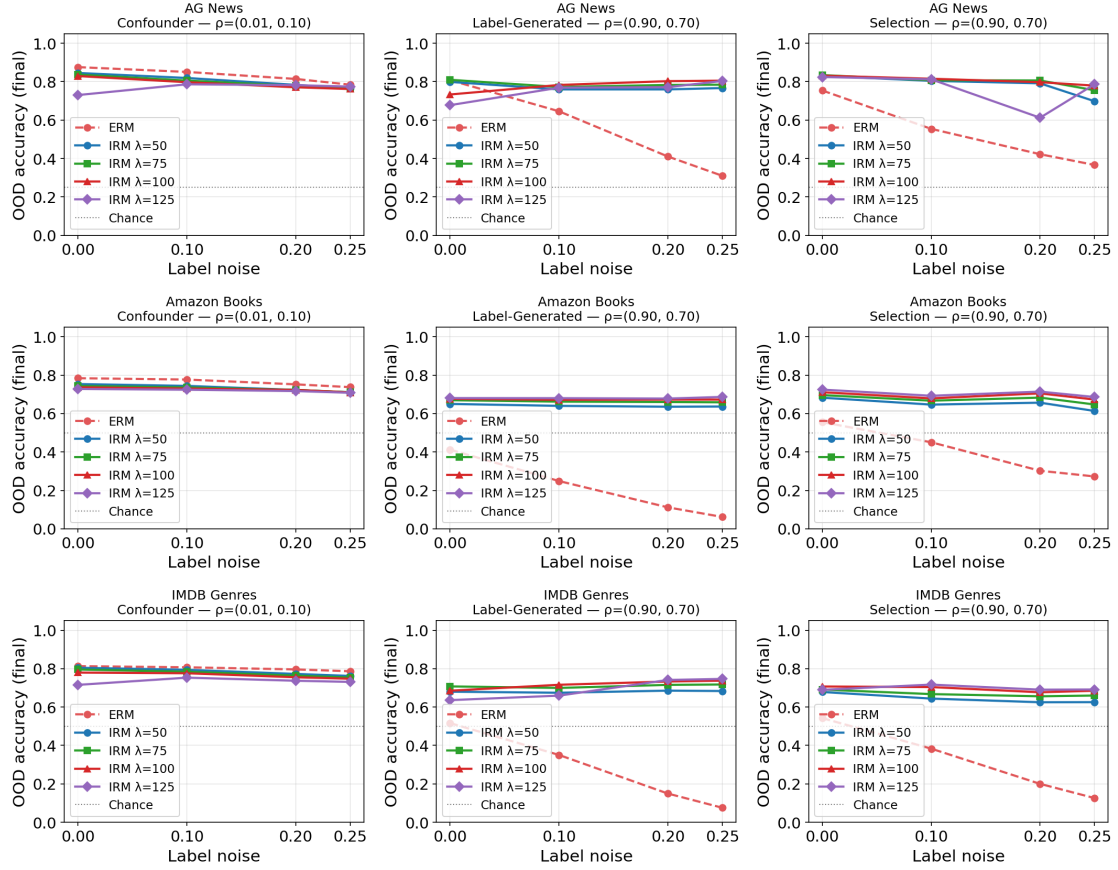


Figure 7: OOD accuracy as a function of label noise, for each $\lambda \in \{50, 75, 100, 125\}$ (solid colored lines) and ERM (dashed red). Each panel corresponds to one dataset–mechanism pair. All IRM variants consistently outperform ERM. Performance differences across λ values are modest, confirming robustness of the chosen $\lambda = 100$.

tested mechanism rather than a violation of the method’s core assumptions. This structural rigor is fundamental to the analysis in the following section, where we evaluate confounding. By distinguishing between three possible modeling scenarios, we clarify which interventions preserve the target mechanism and why invariance of the causal mechanism need not imply invariance of the optimal predictor when relevant causal parents are latent.

C Confounding and IRM Validity

Unlike the other two processes, the confounder U exerts a causal influence on the label Y . For IRM to be theoretically valid, the influence of causal parents on the target must remain stable, as established in Appendix B. We distinguish three environment-generation strategies. The first two illustrate, respectively, a valid intervention on the shortcut mechanism and an invalid intervention on the target mechanism. The third is also structurally valid under the criterion of Appendix B, but highlights the distinction between invariance of the

causal mechanism and invariance of the optimal predictor when causal parents are latent.

- **Varying proxy:** The link between the unobserved confounder U and the latent concept Z varies across environments, but the $U \rightarrow Y$ relationship remains fixed. This is the intervention used in our main confounding experiments: the causal rule remains fixed, and only the misleading proxy relationship fluctuates. This setup preserves the condition $\mathbb{E}[Y^e \mid \text{Pa}(Y)] = \mathbb{E}[Y \mid \text{Pa}(Y)]$.
- **Varying causal strength:** In this scenario, the direct causal link between the confounder U and the target Y varies. This represents a direct violation of IRM’s assumptions, as the conditional expectation $\mathbb{E}[Y \mid \text{Pa}(Y)]$ is no longer invariant. Since no stable relationship exists across environments, IRM systematically fails to identify a robust predictor.
- **Confounder-prevalence shift:** A third possibility is to keep both structural mechanisms

Mechanism	Model	AG News				Amazon Books				IMDB Genres			
		Label-noise rate ϵ				Label-noise rate ϵ				Label-noise rate ϵ			
		0.00	0.10	0.20	0.25	0.00	0.10	0.20	0.25	0.00	0.10	0.20	0.25
Confounding	ERM	0.875	0.851	0.814	0.784	0.784	0.777	0.752	0.738	0.813	0.807	0.796	0.786
	IRM ($\lambda = 50$)	0.845	0.819	0.783	0.765	0.753	0.744	0.723	0.710	0.804	0.794	0.772	0.762
	IRM ($\lambda = 75$)	0.837	0.805	0.778	0.763	0.744	0.736	0.723	0.711	0.795	0.784	0.763	0.753
	IRM ($\lambda = 100$)	0.829	0.797	0.770	0.761	0.737	0.730	0.723	0.710	0.779	0.776	0.755	0.748
	IRM ($\lambda = 125$)	0.730	0.786	0.781	0.774	0.728	0.724	0.717	0.708	0.716	0.753	0.737	0.731
Label-Generated	ERM	0.807	0.645	0.410	0.309	0.412	0.248	0.112	0.061	0.516	0.350	0.149	0.075
	IRM ($\lambda = 50$)	0.798	0.759	0.759	0.766	0.650	0.640	0.635	0.636	0.680	0.674	0.685	0.683
	IRM ($\lambda = 75$)	0.809	0.772	0.782	0.784	0.669	0.662	0.660	0.658	0.708	0.700	0.716	0.718
	IRM ($\lambda = 100$)	0.733	0.782	0.802	0.805	0.676	0.671	0.672	0.672	0.685	0.716	0.733	0.738
	IRM ($\lambda = 125$)	0.677	0.768	0.771	0.803	0.680	0.680	0.678	0.686	0.636	0.659	0.741	0.747
Selection	ERM	0.754	0.554	0.422	0.366	0.555	0.451	0.302	0.273	0.542	0.384	0.200	0.126
	IRM ($\lambda = 50$)	0.830	0.804	0.791	0.698	0.682	0.646	0.656	0.614	0.678	0.644	0.625	0.625
	IRM ($\lambda = 75$)	0.834	0.807	0.806	0.755	0.696	0.667	0.683	0.647	0.691	0.667	0.656	0.660
	IRM ($\lambda = 100$)	0.831	0.815	0.797	0.779	0.711	0.679	0.705	0.673	0.707	0.704	0.677	0.685
	IRM ($\lambda = 125$)	0.824	0.812	0.612	0.786	0.725	0.693	0.714	0.686	0.690	0.717	0.691	0.692

Table 7: Hyperparameter search results. Mean final OOD accuracy for IRM penalty values $\lambda \in \{50, 75, 100, 125\}$ across label-noise rates $\epsilon \in \{0.00, 0.10, 0.20, 0.25\}$. Results are averaged over spurious-correlation levels. Bold indicates the best IRM result for each dataset, mechanism, and noise configuration.

$P(Y | U)$ and $P(Z | U)$ fixed while changing the marginal distribution $P^{(e)}(U)$ across environments. Unlike a change in the $U \rightarrow Y$ mechanism, such a prevalence shift does not by itself violate the validity condition introduced in Appendix B: the causal relationship between Y and its parent U remains unchanged, and therefore $\mathbb{E}[Y^e | U] = \mathbb{E}[Y | U]$. However, because U is unobserved, changing its prevalence can alter predictive conditionals involving observed proxies. In particular, $P^{(e)}(Y | Z) = \sum_u P(Y | u)P^{(e)}(u | Z)$, and $P^{(e)}(u | Z)$ generally depends on $P^{(e)}(U)$. Consequently, a prevalence shift may change the optimal predictor based on the observed representation even though the underlying causal mechanism remains invariant. Whether IRM can recover an invariant predictor then depends on whether the observed representation contains sufficient stable information about the target independently of the prevalence-sensitive proxy.

D The Role of Causal Direction

IRM seeks a representation for which the same predictor remains optimal across environments. However, its formal guarantees rely on a connection between invariance and causality (Arjovsky, 2020; Arjovsky et al., 2019). The implications of this connection depend on the generative direction of the prediction task: causal learning ($X \rightarrow Y$) or anti-causal learning ($Y \rightarrow X$). This distinction can be understood through the Principle of Independent Causal Mechanisms (ICM) (Schölkopf et al., 2012; Peters et al., 2017). The ICM principle states that a

causal generative process consists of autonomous mechanisms: the conditional distribution of an effect given its cause is structurally independent of the marginal distribution of that cause.

D.1 Causal Learning

In a causal prediction task ($X \rightarrow Y$), such as topic classification or review-helpfulness prediction, the observed features X causally determine the target Y . Under the ICM principle, the mechanism $P(Y | X)$ may remain stable even when the distribution of inputs $P^{(e)}(X)$ changes across environments. The joint distribution in environment e can be factorized as

$$P^{(e)}(X, Y) = P^{(e)}(X)P(Y | X), \quad (6)$$

where environmental variation affects $P^{(e)}(X)$ but leaves the causal mechanism $P(Y | X)$ unchanged. In this setting, identifying an invariant representation amounts to preserving the direct causal parents of Y , which define a stable and optimal predictor across valid interventions (Arjovsky, 2020; Arjovsky et al., 2019). IRM is therefore theoretically well aligned with causal prediction when the environmental changes do not modify the mechanism generating Y from its causes.

D.2 Anti-Causal Learning

In an anti-causal prediction task ($Y \rightarrow X$), the target variable generates the observed features. Sentiment classification provides a typical example: the latent sentiment Y generates the vocabulary, expressions, and other textual features contained in X . Predicting Y from X therefore amounts to

inferring a cause from its effects (Jin et al., 2021; Feder et al., 2022). Under the ICM principle, the stable mechanism is now the generative distribution $P(X | Y)$ rather than the predictive distribution $P(Y | X)$. The environment-specific joint distribution factorizes as

$$P^{(e)}(X, Y) = P(X | Y)P^{(e)}(Y), \quad (7)$$

where $P(X | Y)$ is assumed to remain stable, but the class prevalence $P^{(e)}(Y)$ may vary across environments. A classifier estimates the posterior distribution $P^{(e)}(Y | X)$, which follows from Bayes’ theorem:

$$\begin{aligned} P^{(e)}(Y | X) &= \frac{P(X | Y)P^{(e)}(Y)}{P^{(e)}(X)} \\ &= \frac{P(X | Y)P^{(e)}(Y)}{\sum_{y' \in \mathcal{Y}} P(X | Y = y')P^{(e)}(Y = y')} \end{aligned} \quad (8)$$

where $\mathcal{Y}$ denotes the set of possible target labels. Even if the generative mechanism $P(X | Y)$ remains invariant, the optimal predictive distribution $P^{(e)}(Y | X)$ depends explicitly on the environment-specific class prevalence $P^{(e)}(Y)$. This creates two distinct regimes for invariant learning:

- **Stable class prevalence.** If $P^{(e)}(Y)$ is identical across training and test environments, and $P(X | Y)$ remains stable, then $P^{(e)}(Y | X)$ is also invariant. IRM can suppress unstable environmental shortcuts while retaining predictive relationships shared across environments.
- **Label shift.** If $P^{(e)}(Y)$ changes across environments, then the optimal posterior $P^{(e)}(Y | X)$ generally changes even though $P(X | Y)$ remains stable. Enforcing a common optimal predictor across environments then conflicts with the environment-specific predictor required by Bayes’ rule.

Label shift therefore creates a structural difficulty for IRM in anti-causal tasks. The problem is not that the underlying generative mechanism becomes unstable, but that the posterior used for prediction depends on a changing class prior. A strict invariance constraint may consequently prevent the classifier from adapting to the posterior required in each environment.

D.3 Implications for Our Experiments

Our anti-causal experiments are conducted with balanced and constant class proportions across training and test environments. This construction keeps $P^{(e)}(Y)$ fixed while varying only the association between the target and the designated shortcut. Under the assumed stability of $P(X | Y)$, the optimal predictive mechanism can therefore remain invariant across environments. This design isolates IRM’s ability to reject unstable shortcuts without simultaneously introducing label shift. The resulting performance should consequently be interpreted as evidence that IRM can operate in an anti-causal setting under stable class prevalence, rather than as evidence that it is generally robust to arbitrary distribution shifts in anti-causal NLP tasks.

E Methodological Justification of Sequence Length as a Spurious Feature

To ensure that our experimental design in the IMDB Genres setting is theoretically justified, it is necessary to establish that text description length is a feature that models can readily extract and exploit as a predictive shortcut. Linguistic probing and model analysis literature establish that sequence length information is preserved within sentence-level vector spaces (Adi et al., 2017). Consequently, when sentence representations are computed via token averaging (mean pooling), the resulting vector is directly bound to the sequence length N , as embedding norms predictably decrease as sentences grow longer due to the properties of natural language distributions. Transformer-based architectures are highly susceptible to the sequence length learning problem whenever downstream classes exhibit asymmetrical length distributions (Bailargeon and Lamontagne, 2022). In these unbalanced scenarios, standard ERM objectives tend to opportunistically exploit sequence length as a primary predictive feature instead of relying on genuine textual or semantic information. Furthermore, because sequence length distributions effectively behave like a distinct domain or environmental property during training, a significant mismatch or inversion of these length patterns at test time routinely triggers an overfitting collapse (Variš and Bojar, 2021). Modulating this engineered length bias across training environments therefore provides a theoretically justified and rigorous proxy for evaluating the capacity of IRM to bypass surface heuristics in favor of invariant features.

F Detailed Construction of Spurious Mechanisms

To rigorously evaluate IRM across diverse causal structures, we artificially inject or induce spurious correlations in each of the three datasets. Each corpus is partitioned into an 80% training set, a 10% in-distribution validation set, and a 10% out-of-distribution test set. The training partition is further divided equally into two environments $\mathcal{E}_{\text{tr}} = \{e_1, e_2\}$, characterized by contrasting spurious correlation strengths, so as to create the distributional instability required by the invariance penalty. We detail below the exact construction procedure for each of the three mechanisms.

Label-Generated Features (p_c). The target label Y causally generates a discrete textual artifact, which is embedded in the raw text prior to encoding. For each dataset, we associate a set of semantically neutral vocabulary tokens to the possible classes: {moon, sun} for Amazon Books, {red, blue, green, yellow} for AG News (one per class), and {pine, ash} for IMDB Genres. Rather than simply prepending a single token at the start of the text, the selected token is inserted immediately before every occurrence of a predefined list of high-frequency, content-free function words (e.g., *the, a, of, in, to*). This distributes the spurious signal across the entire sequence, making the shortcut substantially harder for ERM to ignore.

The correlation strength is controlled by $p_c = P(Z = y \mid Y = y)$. To prevent the true semantic signal from trivially dominating, a uniform label noise $\epsilon = 0.2$ is applied globally to the entire training partition before splitting into environments, ensuring that both environments share the same noise pattern. The two training environments are then constructed by injecting tokens with $p_c \in \{0.70, 0.90\}$ respectively. In the OOD test set, this correlation is completely inverted ($p_c = 0.0$), so the injected token systematically identifies the wrong class.

Confounding (a). An unobserved confounder C introduces a spurious backdoor path to the observed label. For each training example, C is drawn uniformly from all possible classes for a fraction $p_{c\text{-flip}} = 0.50$ of the data; for the remainder, C coincides with the true label Y^* (inactive confounder). The observed label Y_{obs} is then biased toward C : when $Y^* \neq C$, it is replaced by C with probability γ (set to 0.8 for Amazon and AG News, 0.5

for IMDB Genres). The model never observes C directly; instead, it observes a noisy proxy Z , generated as $Z^{(e)} = C \oplus \text{Ber}(a^{(e)})$, where $\oplus$ is the XOR operator and $a^{(e)}$ governs how faithfully Z tracks C . The injected token encodes Z rather than Y_{obs} , so the spurious correlation propagates through the backdoor path $Y_{\text{obs}} \leftarrow C \rightarrow Z \rightarrow X_Y^\perp$. The two training environments differ solely in the proxy noise rate: $a^{(e)} \in \{0.01, 0.10\}$, keeping Z a reliable proxy in both environments. In the OOD test set, a is set to 1.0, making Z perfectly anti-correlated with C ; evaluation is performed on the clean labels Y^* (without label bias), maximally penalizing any model that has relied on the proxy token.

Selection Bias (p_s). Unlike the token injection mechanisms above, selection bias operates purely by subsampling the corpus based on an unobserved collider S . A pre-existing latent feature Z is defined for each dataset, and a sample is included in training ($S = 1$) based on whether its (Y, Z) pair is typical (aligned) or atypical (misaligned). The parameter p_s controls the proportion of typical examples retained in each environment ($p_s \in \{0.70, 0.90\}$). The OOD test set contains exclusively atypical examples, completely inverting the spurious alignment.

For Amazon Books, the latent feature Z is derived from the star rating: reviews with 4–5 stars are considered positive ($Z = 1$) and reviews with 1–2 stars negative ($Z = 0$). A typical example is one where Z aligns with Y : helpful reviews are positive, unhelpful reviews are negative. The OOD set consists exclusively of counter-heuristic instances (helpful reviews with negative sentiment, and vice versa).

For AG News, Z is the character length of the article, discretized into four global length bins via corpus-wide quartiles (Q1: $\leq P_{25}$, Q2: $P_{25}-P_{50}$, Q3: $P_{50}-P_{75}$, Q4: $> P_{75}$). Each class is assigned a dedicated typical bin: World $\rightarrow$ Q1, Sports $\rightarrow$ Q2, Business $\rightarrow$ Q3, Sci/Tech $\rightarrow$ Q4. In the OOD test set, each class is evaluated exclusively on articles from its most distant bin (i.e., Q1 $\leftrightarrow$ Q4 and Q2 $\leftrightarrow$ Q4 or Q1), ensuring the length cue is maximally misleading at test time.

For IMDB Genres, Z is the description length, thresholded at the global P25 and P75 percentiles to obtain a short ($< P_{25}$) and a long ($> P_{75}$) category; intermediate-length texts are excluded. A typical example is a long Romance or a short

Thriller. Training environments are assembled from $p_s \in \{0.70, 0.90\}$ typical pairs, while the OOD set contains exclusively atypical instances: short Romances and long Thrillers.

G Additional Bias in Bios Details

This section provides supplementary details on the construction of the Bias in Bios experiments introduced in Section 6.1, including the environment-generation procedure, the semantic-distance analysis used to select profession pairs, and an additional multiclass experiment.

G.1 Dataset and Environment Construction

Bias in Bios contains approximately 400,000 short biographies annotated with one of 28 professions and binary gender (De-Arteaga et al., 2019). We construct binary classification tasks from pairs containing one male-typical and one female-typical profession. A profession is defined as male-typical when fewer than 50% of its biographies are annotated as female, and female-typical otherwise. Let Y denote the profession label, G the gender annotation, and $\text{canonical}(Y)$ the majority gender of profession Y in the original corpus. Shortcut alignment is defined as

$$A = \mathbf{1}[G = \text{canonical}(Y)]. \quad (9)$$

Gender is not provided as a separate input, but may be inferred from names, pronouns, and other gender-associated expressions. Two training environments are produced by resampling biographies within each profession while preserving class balance. Their shortcut strengths are

$$\begin{aligned} P^{(e_1)}(A = 1 \mid Y) &= 0.99, \\ P^{(e_2)}(A = 1 \mid Y) &= 0.80. \end{aligned} \quad (10)$$

The OOD test environment fully reverses the association, $P^{(\text{test})}(A = 1 \mid Y) = 0$, so that every test biography belongs to the minority-gender subset of its profession. A classifier relying on the training-time gender association therefore receives a systematically misleading cue at test time. Because profession-specific vocabulary is highly predictive, we apply a label-flip rate of $\epsilon = 0.25$ to the training labels to make the gender cue competitive. Gender annotations and biography texts are left unchanged, and evaluation uses the original profession labels. We use the frozen DistilBERT encoder and masked mean pooling described in Appendix A. Only the

linear classification head is optimized. ERM pools the two environments, whereas IRMv1 retains the environment partition and uses a penalty coefficient of $\lambda = 1000$. Both models are trained for 3,000 steps with Adam and a learning rate of 10^{-4} .

G.2 Semantic Distance and Pair Selection

We quantify profession similarity in the representation space of the same frozen encoder. For each profession c , up to 300 biographies are sampled from the full corpus and encoded using masked mean pooling. The normalized profession centroid is

$$\tilde{\phi}_c = \frac{\frac{1}{n_c} \sum_{i:Y_i=c} \Phi_{\text{BERT}}(x_i)}{\left\| \frac{1}{n_c} \sum_{i:Y_i=c} \Phi_{\text{BERT}}(x_i) \right\|_2}. \quad (11)$$

For a male-typical profession c_m and a female-typical profession c_f , their distance is

$$d(c_m, c_f) = 1 - \tilde{\phi}_{c_m}^\top \tilde{\phi}_{c_f}. \quad (12)$$

After applying the corpus-availability criteria used in preprocessing, 160 cross-gender profession pairs are retained and ranked by this distance. Lower values indicate greater similarity in the frozen representation space. For the experiments in Section 6.1, we select five semantically close and five semantically distant eligible pairs while requiring enough examples for repeated balanced resampling with at least 1,000 examples per training environment. The ten evaluated pairs and their OOD results are reported in Table 4. The close group is attorney/paralegal, professor/psychologist, chiropractor/psychologist, chiropractor/nurse, and physician/nurse; the distant group is dj/nurse, rapper/psychologist, dj/paralegal, rapper/paralegal, and rapper/nurse.

Semantic distance is used only as a descriptive proxy for the availability of profession-specific information. It does not directly measure the accessibility of gender, the frequency of explicit gender markers, within-class dispersion, or the orientation of the learned decision boundary. Consequently, distance should not be expected to predict ERM shortcut reliance or the IRM gain monotonically. Its role is to construct comparisons with different levels of representation-space separability, rather than to provide a complete model of shortcut attractiveness. For reference, the closest pair in the retained ranking is attorney/paralegal, with distance 0.0094, while the maximum observed distance is 0.1489 for rapper/nurse. The complete ranking

is provided with the accompanying experimental material.

G.3 Supplementary Multiclass Experiment

We additionally construct a six-class task to test whether the shortcut remains competitive when the prediction space contains professions from several distinct domains. Training environments are created using the same resampling principle, and the gender–profession association is inverted at test time. In this single-seed experiment, ERM obtains an ID validation accuracy of 0.897 and an OOD accuracy of 0.706, compared with a chance level of 1/6. Only 25.6% of its predictions follow the inverted gender cue. These results suggest that gender is less competitive when profession-specific content provides strong class separation. This experiment remains supplementary because it uses a single seed and is not directly comparable to the binary tasks reported in Table 4.

G.3.1 Model and Optimization

We use the same frozen DistilBERT encoder and linear classification head described in Appendix A. Biography representations are obtained using masked mean pooling over the final hidden states. Only the linear classification head is optimized. ERM pools the two training environments and minimizes the standard classification loss. IRMv1 retains the environment partition and optimizes the sum of the environment-specific risks and the invariance penalty. Both models are trained for 3,000 gradient steps with Adam and learning rate 10^{-4} . For this experiment, the IRMv1 penalty coefficient is set to $\lambda = 1000$. The larger penalty relative to the controlled experiments reflects the strong profession-specific semantic signal in Bias in Bios and the more diffuse nature of the demographic shortcut. All other model components are shared between ERM and IRM.

H ActorCorr: Experimental Details and Additional Results

H.1 Experimental Setup

ActorCorr is based on IMDB movie reviews and the actor-name shortcut protocol of Eshuijs et al. (2025). We use the authors’ data-generation pipeline, which identifies existing actor mentions through named-entity recognition and replaces an eligible mention at a fixed injection frequency of 5%. Following the original pairing, Morgan

Freeman is associated with positive sentiment and Adam Sandler with negative sentiment. Let Z denote the injected actor and Y the sentiment label. Shortcut purity in environment e is defined as

$$\pi^{(e)} = P^{(e)}(Z = z_{\text{canonical}}(Y) \mid \text{actor injected}), \quad (13)$$

where $z_{\text{canonical}}(Y)$ is the actor assigned to class Y . We construct two training environments with

$$\pi^{(e_1)} = 1.0, \quad \pi^{(e_2)} = 0.85. \quad (14)$$

The injection frequency is identical across environments, and only the actor–label agreement varies. Unlike the neutral tokens used in the controlled experiments, actor names may carry associations inherited from pretraining. ActorCorr therefore tests IRM against a semantically loaded shortcut. We use the same frozen DistilBERT encoder and trainable linear classification head as in the controlled experiments. ERM pools both environments, whereas IRMv1 retains the environment partition when computing its risk and invariance penalty. Models are trained for 25,000 gradient steps, and results are averaged over three random seeds. The main ActorCorr experiment uses $\lambda = 100$. Each test review is evaluated under three conditions:

- **Clean:** no actor cue is injected;
- **Correlated:** the actor agrees with the sentiment label;
- **Anti-correlated:** the actor is associated with the opposite label.

To measure shortcut sensitivity independently of pooled accuracy, we compare each review with its anti-correlated variant, in which the injected actor cue points toward the incorrect sentiment class. For each sentiment class $y \in \mathcal{Y}$, we define the class-specific clean-to-anti-correlated accuracy gap as

$$G_y = \text{Acc}_{\text{clean}}^{(y)} - \text{Acc}_{\text{anti}}^{(y)}, \quad (15)$$

where $\text{Acc}_{\text{clean}}^{(y)}$ is the accuracy on clean reviews from class y , and $\text{Acc}_{\text{anti}}^{(y)}$ is the accuracy on the same class under the anti-correlated actor intervention. Since ActorCorr is a binary sentiment-classification task with $\mathcal{Y} = \{\text{neg}, \text{pos}\}$, this definition yields two class-specific gaps:

$$G_{\text{neg}} = \text{Acc}_{\text{clean}}^{(\text{neg})} - \text{Acc}_{\text{anti}}^{(\text{neg})}, \quad (16)$$

$$G_{\text{pos}} = \text{Acc}_{\text{clean}}^{(\text{pos})} - \text{Acc}_{\text{anti}}^{(\text{pos})}. \quad (17)$$

Here, G_{neg} measures the accuracy loss caused by injecting the positive-associated actor into negative reviews, whereas G_{pos} measures the loss caused by injecting the negative-associated actor into positive reviews. These gaps therefore quantify how strongly the misleading actor cue affects each sentiment class. A large positive value of G_y indicates that the anti-correlated actor reduces accuracy, showing that the model remains sensitive to the shortcut for class y . A value close to zero indicates that the intervention has little effect. A negative value means that accuracy is higher under the anti-correlated intervention than in the clean condition, indicating a sign-inverted response rather than straightforward invariance. We summarize the model’s average signed sensitivity using the Anti-Correlated Accuracy Change:

$$\text{ACAC} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} G_y. \quad (18)$$

Because ActorCorr contains two sentiment classes, this expression reduces to

$$\text{ACAC} = \frac{G_{\text{neg}} + G_{\text{pos}}}{2}. \quad (19)$$

A large positive ACAC indicates that anti-correlated actor cues reduce accuracy on average, suggesting reliance on the shortcut. An ACAC close to zero indicates limited average signed sensitivity. However, reporting G_{neg} and G_{pos} in addition to ACAC is important because the effect of the actor cue may be asymmetric across sentiment classes. In particular, a positive gap for one class and a negative gap for the other can partially cancel in the macro-average, producing a low ACAC even though the model remains sensitive to the intervention. Consequently, a robust model should ideally combine high clean and anti-correlated accuracy with both G_{neg} and G_{pos} close to zero. The aggregate ACAC measures average signed sensitivity, while the two class-specific gaps show whether that apparent invariance holds consistently across both sentiment classes.

H.2 Results by Test Condition

Table 8 decomposes the pooled OOD results by test condition.

ERM strongly benefits from the correlated actor cue but deteriorates when the cue contradicts the label, producing a spread of 0.296 across test conditions. IRM is less accurate on clean and correlated

Test condition	ERM	IRM
Clean	0.759	0.716
Correlated	0.870	0.726
Anti-correlated	0.574	0.671
Spread	0.296	0.055

Table 8: ActorCorr accuracy by test condition, averaged over three seeds. Spread denotes the difference between the highest and lowest condition-level accuracies.

λ	OOD	ACAC	G_{neg}	G_{pos}
10	0.745 ± 0.007	0.146 ± 0.002	0.179 ± 0.001	0.114 ± 0.002
25	0.740 ± 0.005	0.120 ± 0.002	0.160 ± 0.003	0.080 ± 0.002
50	0.732 ± 0.004	0.082 ± 0.003	0.134 ± 0.004	0.031 ± 0.001
75	0.726 ± 0.008	0.061 ± 0.000	0.124 ± 0.002	-0.001 ± 0.002
100	0.704 ± 0.008	0.049 ± 0.002	0.148 ± 0.002	-0.051 ± 0.003
ERM	0.739 ± 0.004	0.186 ± 0.002	0.221 ± 0.001	0.151 ± 0.002

Table 9: ActorCorr sensitivity to λ , reported as mean $\pm$ standard deviation over three seeds. G_y is the clean-to-anti-correlated accuracy gap for class y ; negative values indicate a sign-inverted response.

examples, but substantially more accurate under the anti-correlated intervention. Its spread is reduced to 0.055, indicating more homogeneous behavior across conditions. Pooled OOD accuracy therefore understates the difference in shortcut reliance, since ERM’s gains on correlated examples partially compensate for its failures on anti-correlated examples.

H.3 Sensitivity to the Invariance Penalty

Because ActorCorr combines a relatively small purity difference across environments with semantically meaningful actor names, we evaluate the following penalty values:

$$\lambda \in \{10, 25, 50, 75, 100\}. \quad (20)$$

All configurations use the same architecture, training environments, and optimization protocol.

Increasing λ monotonically reduces ACAC from 0.146 ± 0.002 to 0.049 ± 0.002 , but also lowers pooled OOD accuracy from 0.745 ± 0.007 to 0.704 ± 0.008 . ActorCorr therefore reveals a clear accuracy–invariance trade-off: the penalty value that minimizes sensitivity to the targeted shortcut does not maximize aggregate predictive performance. Relative to the ERM reference, every IRM configuration reduces ACAC and both class-specific gaps, confirming that the invariance penalty consistently weakens reliance on the actor cue. However, only $\lambda \in \{10, 25\}$ matches or

Injection	Model	OOD	ACAC	G_{neg}	G_{pos}
10%	ERM	0.732 ± 0.007	0.160 ± 0.005	0.203 ± 0.011	0.118 ± 0.003
	IRM, 10	0.741 ± 0.008	0.124 ± 0.002	0.174 ± 0.001	0.075 ± 0.003
	IRM, 25	0.739 ± 0.005	0.100 ± 0.001	0.138 ± 0.003	0.062 ± 0.001
	IRM, 50	0.734 ± 0.005	0.071 ± 0.003	0.115 ± 0.003	0.027 ± 0.002
	IRM, 75	0.725 ± 0.005	0.055 ± 0.003	0.115 ± 0.003	-0.004 ± 0.004
	IRM, 100	0.709 ± 0.005	0.043 ± 0.000	0.142 ± 0.002	-0.055 ± 0.003
25%	ERM	0.717 ± 0.008	0.291 ± 0.003	0.324 ± 0.006	0.258 ± 0.003
	IRM, 10	0.724 ± 0.008	0.227 ± 0.001	0.256 ± 0.004	0.198 ± 0.002
	IRM, 25	0.728 ± 0.009	0.170 ± 0.003	0.202 ± 0.006	0.137 ± 0.003
	IRM, 50	0.728 ± 0.008	0.115 ± 0.004	0.164 ± 0.011	0.065 ± 0.008
	IRM, 75	0.716 ± 0.009	0.098 ± 0.005	0.162 ± 0.002	0.033 ± 0.009
	IRM, 100	0.697 ± 0.007	0.071 ± 0.004	0.176 ± 0.013	-0.034 ± 0.005

Table 10: ActorCorr sensitivity to effective actor-injection frequency and IRM penalty, reported as mean $\pm$ standard deviation over three seeds. G_y is the clean-to-anti-correlated accuracy gap for class y ; negative values indicate a sign-inverted response.

improves upon ERM’s OOD accuracy, whereas stronger penalties achieve greater invariance at the cost of predictive performance. The class-specific gaps further show that this trade-off is asymmetric. For the negative class, G_{neg} remains positive for all penalty values. For the positive class, the gap approaches zero at $\lambda = 75$ and becomes negative at $\lambda = 100$, with $G_{\text{pos}} = -0.051 \pm 0.003$. The strongest penalty therefore induces a sign-inverted response for positive examples rather than simply removing sensitivity to the actor cue. This asymmetry may result from unequal associations of the two actor names in the frozen pretrained representation space. Consequently, a single global penalty can suppress the two class-specific shortcut responses at different rates.

H.4 Sensitivity to Shortcut Prevalence

The main ActorCorr experiment fixes the effective actor-injection frequency at 5%. To assess how the prevalence of the shortcut changes the comparison, we additionally generate two datasets in which 10% or 25% of reviews in each training environment receive an actor substitution. We retain the same actor-label purities, $\pi^{(e_1)} = 1.0$ and $\pi^{(e_2)} = 0.85$, the same three evaluation conditions, and the same optimization protocol. Thus, this intervention changes the prevalence of the actor cue without changing its environment-specific agreement mechanism.

Table 10 shows that the frequency manipulation successfully controls shortcut attractiveness. Increasing the injection frequency from 10% to 25% raises ERM’s ACAC from 0.160 ± 0.005 to 0.291 ± 0.003 , while lowering its pooled OOD accuracy from 0.732 ± 0.007 to 0.717 ± 0.008 . At

25%, IRMv1 with $\lambda = 50$ is the best accuracy–invariance compromise: it improves OOD accuracy to 0.728 ± 0.008 and reduces ACAC to 0.115 ± 0.004 . This corresponds to a 60.5% reduction in signed average sensitivity relative to ERM, while both class-specific gaps remain positive and smaller than their ERM counterparts.

The lowest ACAC values are obtained with $\lambda = 100$, but this setting also reduces OOD accuracy and makes G_{pos} negative at both injection frequencies. As in the fixed-frequency sweep, this behavior is evidence of a sign-inverted response for the positive class rather than complete invariance. At the weaker 10% prevalence, the OOD optimum occurs at $\lambda = 10$; larger penalties trade accuracy for lower shortcut sensitivity. The frequency sweep therefore confirms that the useful penalty scale depends on shortcut prevalence, while preserving the same qualitative accuracy–invariance trade-off observed in Table 9.