Mitigating Spurious Correlations in Text Classification Using Latent Space Geometry

Jiasen Gao[1], **Xiaoliang Chen**[1,2,4*], **Duoqian Miao**[2], **Xu Gu**[3], **Xianyong Li**[1] and **Yajun Du**[1]
[1]Xihua University, Chengdu, China
[2]Tongji University, Shanghai, China
[3]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu, China
[4]University of Montreal
chenxl@mail.xhu.edu.cn

Abstract

Spurious correlations cause deep learning models to rely on predictive shortcuts that hold in the training data but break under distribution shifts, leading to large performance drops for minority groups. Existing strategies often rely on costly group annotations or employ unstable adversarial training. In this paper, we propose Prototype-guided debiasing using Robust Invariant Feature Transformations (PRIFT), a novel framework that mitigates spurious correlations by manipulating latent space geometry. Specifically, we introduce a prototype-guided modeling approach that leverages natural language prompts to represent confounders, transforming abstract biases into interpretable geometric anchors without auxiliary classifiers. Based on these anchors, we introduce a centered projection operator that adaptively purifies representations by removing confounding deviations specific to instances while preserving essential semantic structure. Furthermore, PRIFT can handle confounding factor information at different levels, ranging from true labels to unsupervised latent inference. Experiments on four text classification benchmarks demonstrate the superiority of our method; notably, PRIFT outperforms state-of-the-art baselines and improves worst-group accuracy by over 20% on the CivilComments dataset compared to standard empirical risk minimization.

1 Introduction

Text classification models have achieved remarkable performance across various natural language processing tasks, yet their reliability is often compromised by spurious correlations, which are predictive shortcuts that hold in the training data but fail under out-of-distribution (OOD) settings (Aldrich, 1995; Geirhos et al., 2020; Gururangan et al., 2018; Hendrycks et al., 2020; Sagawa et al., 2019). In the context of natural language processing (NLP), linguistic style, author demographics, or dataset artifacts can unintentionally serve as such shortcuts, causing models to rely on superficial patterns rather than semantic content. Consider the Yelp-Author-Style dataset (Zhou et al., 2024) shown in Figure 1. Although both reviews convey negative sentiment, classifiers tend to erroneously correlate Shakespearean complex syntactic structures with positive sentiment. In contrast, an ideal classifier should make predictions based on the actual semantics of the reviews. When such spurious correlations do not hold in the real world, classifier performance degrades.

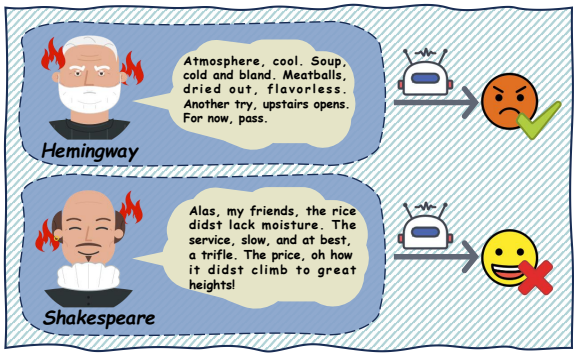


Figure 1: This example shows that the model's misclassification is caused by different writing styles (Shakespeare and Hemingway).

Prior studies divide data into groups based on combinations of class labels and spurious features. When the sample sizes across these groups are imbalanced, classifiers trained on such data tend to rely on spurious correlations. This results in substantially worse performance for minority groups than for majority groups. Approaches like Group Distributionally Robust Optimization

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27525–27539

(Group DRO) (Sagawa et al., 2019) explicitly optimize for Worst-Group Accuracy (WGA) (Liu et al., 2021; Goel et al., 2021) but require expensive, fine-grained group labels, which are often impractical to obtain. On the other hand, geometric methods such as Iterative Null-space Projection (INLP) (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023) operate by identifying and removing subspaces corresponding to spurious attributes. However, such methods usually adopt "hard" orthogonal projection operations, which require training auxiliary classifiers. These operations suffer significant degradation when spurious attributes overlap with task semantics. In this paper, we propose Prototype-guided debiasing using Robust Invariant Feature Transformations (PRIFT), a framework that mitigates spurious correlations by manipulating latent space geometry through natural language prompts. Instead of training auxiliary probes, we leverage the encoder's semantic understanding, encoding a textual prompt describing the bias to generate a geometric anchor directly in the embedding space. While large language models (LLMs) show promise in zero-shot reasoning, our analysis reveals that PRIFT enables standard encoders to achieve superior group robustness with significantly lower computational footprints.

Unlike standard hard projections that completely eliminate information, we design a novel centered projection operator. It removes only the deviations from the group mean along the confounder direction while preserving the global semantic structure. This enables the model to ignore variations in spurious features without compromising the integrity of the underlying representations. We integrate this mechanism into a unified framework that supports varying levels of supervision, from ground truth labels to unsupervised latent inference. This gives our geometric debiasing method a "plug-and-play" property. The main contributions of this work are summarized as follows:

1. We propose a prototype-guided modeling paradigm that transforms abstract biases into interpretable geometric anchors using natural language prompts, eliminating the need for auxiliary classifiers.

2. We introduce a centered projection operator that adaptively purifies representations. By anchoring projections to the group-wise mean, we mitigate spurious variance while preserving essential semantic information.

3. Compared with standard baselines on four text classification tasks with varying spurious correlations, PRIFT achieves better performance on most tasks and even outperforms methods that rely on full group labels.

2 Related work

2.1 Robustness and bias

A significant research effort (Sagawa et al., 2019; Nguyen et al., 2023) improves robustness under OOD settings by explicitly optimizing performance on underrepresented subgroups. Group DRO upweights high loss groups to prevent the model from neglecting minority subpopulations. However, these methods assume access to group labels defined by combinations of target labels and spurious attributes, which are expensive and often impractical to obtain. Another line of research seeks to learn invariant representations across environments, e.g., via invariant risk minimization (IRM) (Wang et al., 2025) and its relaxations (Krueger et al., 2021; Creager et al., 2021). While promising, such methods can be sensitive to how environments are defined or inferred, and may suffer from training instability in practice. In NLP, fairness-oriented debiasing has investigated the ability to remove demographic information from representations (Rakshit et al., 2025; Elsafoury et al., 2023). However, robustness to spurious correlations is broader than demographic fairness. The goal is to eliminate reliance on any irrelevant feature that correlates with labels in the training data (Kumar et al., 2023; Ravfogel et al., 2020).

2.2 Spurious correlation mitigation

Specific techniques have been developed to directly disrupt spurious associations. Causal interventions, such as Counterfactually Augmented Data (CAD) (Kaushik et al., 2020), mitigate bias by editing texts to alter causal features while keeping spurious attributes constant, and CCR (Zhou and Zhu, 2025) combines causal inference with robust optimization. However, these methods often face scalability bottlenecks due to data generation costs or strong causal assumptions. Alternatively, geo-

metric concept erasure methods operate directly on the latent space. INLP (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023) identify confounder subspaces using trained linear probes and project representations onto their null space. However, these methods typically enforce a "hard" orthogonal projection that removes all variance along the confounder direction, potentially damaging semantic content if the attribute partially overlaps with the target label. Furthermore, they rely on auxiliary classifiers to define erasure directions. In contrast, PRIFT leverages natural language prompts to instantiate confounder prototypes from known spurious attributes without training auxiliary classifiers, and employs a centered projection that removes only spurious deviations while preserving the global semantic structure.

3 Method

We propose PRIFT (Figure 2), a framework that mitigates spurious correlations by geometrically purifying latent representations. We assume a dataset where each input x is associated with a target label y and a potential confounder s. Unlike standard settings that require explicit group labels $g = (y, s)$, we treat labels and confounders as natural language prompts.

3.1 Prototype-guided confounder modeling

Both the confounder and the label are represented as natural language prototypes. For each confounder $s \in \mathcal{S}$, a prompt is written to describe the associated attribute (e.g., a particular style or writing pattern). For each label, a label prompt l is written to state the semantic meaning of the label (e.g., positive or negative sentiment). All prompts are encoded with the same text encoder as the review text x. Concretely, the prompt encoder and text encoder share the same architecture and parameters, i.e., $f_{enc} \equiv f_\theta$; we use the two notations only to distinguish their roles when encoding prompts and input texts, respectively.

Formally, given a confounding prompt s and a label prompt l, the corresponding prototype representations are defined as follows:

$$v_s = f_{\text{enc}}(s), \qquad v_l = f_{\text{enc}}(l), \tag{1}$$

where $v_s, v_l \in \mathbb{R}^d$ lie in the same latent space as the raw text feature $h_x = f_\theta(x)$. The encoder output is typically obtained via a pooling operation and then normalized so that these vectors possess a clear geometric interpretation, such as in the context of inner products or cosine similarity measures.

Since prototypes come from readable prompts rather than being learned as free parameters, they can be reused across tasks when the confounder is known conceptually. In practice, we construct these prompts from dataset documentation or evaluation protocols, and updating confounder descriptions reduces to editing the prompts and recomputing the prototypes. This choice also separates confounder and label specification from the encoder architecture, so the same pretrained encoder can be applied across datasets while only the prompts are changed. Group structure in the data is determined jointly by the label prompt and the confounder prompt. Each sample is associated with a pair (l, s), and $g = (l, s)$ denotes the corresponding group identity. The confounder prototype v_s represents characteristics shared by samples with confounder s, while the label prototype v_l represents the semantics associated with the label described by l.

3.2 Adaptive confounder-intensity estimation

Based on these geometric anchors, we quantify the extent to which an input text exhibits the confounding attribute. We define the confounding intensity α of a sample x with respect to a confounder s as the scalar projection of its representation h_x onto the confounder prototype v_s:

$$\alpha = \langle h_x, v_s \rangle \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. A larger α indicates a stronger presence of the confounding signal (e.g., a more intense writing style) in the latent representation.

To distinguish between the inherent semantic structure of the confounder group and instance-specific deviations that lead to spurious correlations, we estimate the stratum-wise mean intensity $\bar{\alpha}_s$ for each confounder s. This is computed by averaging the confounding intensities over all training samples associated with s:

$$\bar{\alpha}_s = \mathbb{E}[\alpha | s] \approx \frac{1}{|D_s|} \sum_{x \in D_s} \langle h_x, v_s \rangle \tag{3}$$

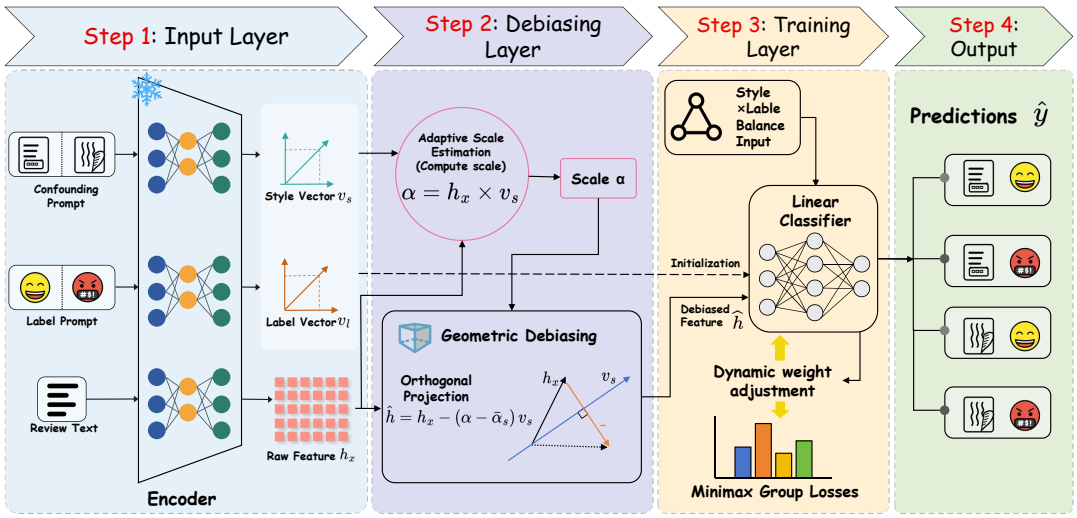


Figure 2: Schematic representation of the proposed PRIFT framework. The processing pipeline comprises four sequential stages: (1) **Input Layer**: Projection of both textual inputs and prompts into a shared latent representation space; (2) **Debiasing Layer**: Estimation of confounding intensity followed by application of a centered projection operator to attenuate or remove confounding information from the representations; (3) **Training Layer**: Optimization of a linear classifier parameterization under a dynamic group reweighting scheme to mitigate group-specific biases; and (4) **Output Layer**: Generation of predictions at the group level. Note that the framework relies on dataset-wise confounder specification via prompts, without strictly requiring instance-wise confounder labels.

where $D_s = x \in D : \gamma(x) = s$ denotes the subset of training data assigned to confounder s under the assignment function $\gamma(\cdot)$. When there exist spurious correlations between confounders and labels, $g = (l, s)$ tend to cluster in distinct regions along the direction of v_s. This clustered separation causes the classifier to rely on shortcut signals induced by confounders to reduce training error, instead of learning robust and semantically meaningful features. PRIFT leverages α to capture the intensity of confounders in individual samples, uses $\bar{\alpha}_s$ as a reference anchor, and the deviation term $(\alpha - \bar{\alpha}_s)$ corresponds to the instance-specific component that we aim to eliminate in the subsequent debiasing step.

3.3 Centered projection debiasing

We propose a centered projection operator to purify the representations. The complete debiasing procedure is outlined in Appendix G. Formally, we derive the debiased feature $\hat{h}$ by subtracting the deviation components along the confounder directions:

$$\hat{h} = h_x - \sum_{j=1}^{k} \lambda_j (\alpha_j - \bar{\alpha}_{s_j}) v_{s_j} \tag{4}$$

where $S = \{s_1, \ldots, s_k\}$ denotes the set of confounders, $\lambda_j > 0$ denotes the debiasing strength for confounder s_j, $\alpha_j = \langle h_x, v_{s_j} \rangle$, and the single-confounder case is recovered by setting $k = 1$. Consequently, $\hat{h}$ retains the average level of confounder information for each confounder group while removing the instance-specific deviations that are most likely to spuriously correlate with the label. This operator satisfies two critical geometric properties that ensure both debiasing effectiveness and semantic preservation.

The centered projection (Eq. 4) calibrates the confounder intensity of all samples within group s to a fixed value. Given $\langle v_s, v_s \rangle = 1$, we derive:

$$\begin{aligned} \langle \hat{h}, v_s \rangle &= \langle h_x, v_s \rangle - (\alpha - \bar{\alpha}_s) \langle v_s, v_s \rangle \\ &= \alpha - (\alpha - \bar{\alpha}_s) \\ &= \bar{\alpha}_s \end{aligned} \tag{5}$$

Consequently, every debiased representation $\hat{h}$ associated with confounder s shares the same projection $\bar{\alpha}_s$ onto v_s. This mechanism effectively neutralizes variance in confounder alignment within each group while maintaining the semantic structure captured by the mean.

Crucially, the operator preserves information in

directions orthogonal to v_s. Let $u \in \mathbb{R}^d$ be any vector such that $\langle u, v_s \rangle = 0$. Then,

$$\langle \hat{h}, u \rangle = \langle h_x, u \rangle - (\alpha - \bar{\alpha}_s)\langle v_s, u \rangle = \langle h_x, u \rangle \quad (6)$$

Consequently, if the label prototype v_l is decomposed into a component aligned with v_s and a component orthogonal to it, the latter is exactly preserved by the centered projection. Since task-relevant semantics are often encoded in directions not purely aligned with the confounder, this property allows the transformation to focus on mitigating variability induced by the confounder without erasing all label information indiscriminately.

Compared to hard projection methods that remove the entire component of h_x along v_s, the centered projection offers a more refined geometric debiasing strategy. By anchoring the projection at the confounder-wise mean $\bar{\alpha}_s$ rather than at zero, it preserves the coarse confounder structure encoded by the encoder while suppressing the fine-grained confounder deviations most responsible for spurious correlations. In the overall architecture, the operation defined in Eq. 4 serves as the confounder debiasing layer, positioned between the encoder and the classifier.

3.4 Unified framework and robust optimization

The PRIFT framework supports different levels of supervision for the confounder variable s, integrating prototype-guided classifier initialization with a robust optimization scheme for groups. Explicitly, our intended supervision regime is dataset-wise confounder specification via a small set of prompts, while optionally employing weak confounder assignment strategies. To accommodate datasets where confounder annotations range from fully available to entirely missing, our framework allows s to be derived via interchangeable mechanisms while maintaining consistent downstream debiasing and training procedures. Given a raw feature h_x and the set of confounder prototypes $\{v_s\}_{s \in \mathcal{S}}$, we determine the confounder assignment s as:

$$s = \begin{cases} s_{gt} & \text{oracle / supervised} \\ \arg\max_{s' \in \mathcal{S}} \langle h_x, v_{s'} \rangle & \text{prototype-based (label-free)} \\ \arg\max_{s' \in \mathcal{S}} [g_\phi(h_x)]_{s'} & \text{probe-based} \end{cases} \quad (7)$$

For the probe-based setting, g_ϕ is defined as a linear softmax probe, $g_\phi(h_x) = \text{Softmax}(W_\phi h_x + b_\phi)$, which is trained with cross-entropy loss exclusively on a subset of data where ground-truth labels s_{gt} are available. Furthermore, we note that the prototype-based $\arg\max$ operation is utilized strictly for geometric anchor selection, rather than calibrated confounder prediction.

Using $\hat{h}$, we train a linear classifier to predict group labels $g = (l, s) \in \mathcal{G} = \mathcal{L} \times \mathcal{S}$. Let $W \in \mathbb{R}^{|\mathcal{G}| \times d}$ denote the classifier weight matrix. To align the classifier with the latent geometry of both labels and confounders, we initialize W from the label and confounder prototypes via:

$$W_0 = \text{Init}(v_l, v_s) \quad (8)$$

where $\text{Init}(\cdot)$ constructs an initial weight vector for each group $g = (l, s)$ from the corresponding pair of prototypes in our implementation, $\text{Init}(v_l, v_s) = \text{norm}(v_l + v_s)$. This prototypically guided initialization yields specific group directions that encode label semantics and confounder characteristics before training on a specific task.

Let L_g denote the empirical loss for group g, computed as the average loss over all samples with group identity $g = (l, s)$. We maintain a group weight vector $q = (q_g)_{g \in \mathcal{G}}$ on the probability simplex to prioritize poorly performing groups during training. Initialized as a uniform distribution, the weights are updated iteratively via an exponential reweighting scheme:

$$q_g \leftarrow q_g \exp(\eta L_g) \quad (9)$$

followed by a normalization step to ensure $\sum_{g \in \mathcal{G}} q_g = 1$, where $\eta > 0$ is a learning-rate parameter controlling the aggressiveness of the reweighting. This update increases the weights of groups with higher losses and decreases those of easier groups, approximating a DRO objective.

In each training iteration, the classifier parameters W are updated to minimize the weighted

average loss:

$$\mathcal{L}(W) = \sum_{g \in \mathcal{G}} q_g L_g(W) \tag{10}$$

where $L_g(W)$ denotes the group-specific loss under the current classifier. By alternating between minimizing $\mathcal{L}(W)$ with respect to W and updating q to focus on high-loss groups, the procedure approximates a minimax problem over classifiers and group distributions. Combined with centered projection debiasing, this training scheme encourages the model to perform well even on groups that would otherwise be disadvantaged by confounder label shortcuts in the training data.

3.5 Output and group level evaluation

After training, the model predicts a probability distribution over group labels for each debiased feature $\hat{h}$ via:

$$\hat{y} = \text{softmax}(W\hat{h}) \tag{11}$$

where each output dimension corresponds to a specific group $g = (l, s) \in \mathcal{G}$. For evaluation, we focus on group level performance metrics that measure robustness against spurious correlations. Let $\text{Acc}(g)$ denote the accuracy on group g, defined as:

$$\text{Acc}(g) = \mathbb{E}\left[\mathbb{1}\left(\hat{y} \text{ correctly predicts } l\right) \,\middle|\, g = (l, s)\right] \tag{12}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The WGA is subsequently defined as:

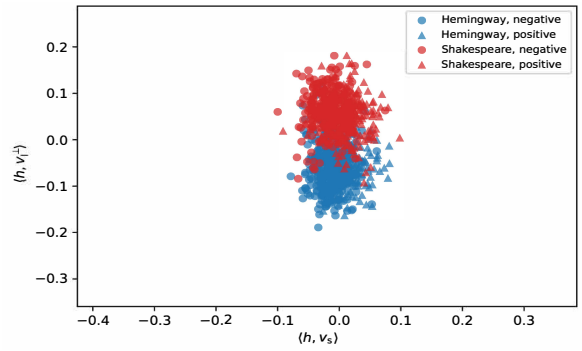
$$\text{WGA} = \min_{g \in \mathcal{G}} \text{Acc}(g) \tag{13}$$

A high WGA indicates that the model maintains strong performance even on the most challenging and potentially underrepresented group, reflecting robustness to spurious correlations induced by confounders.

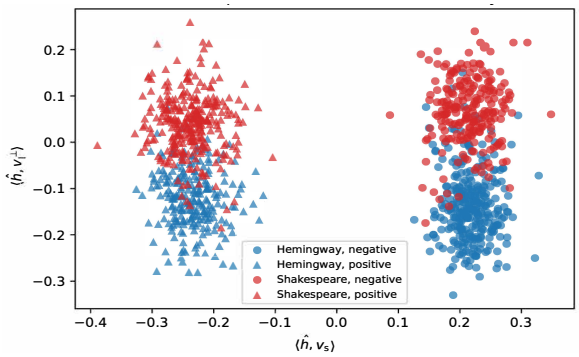
4 Experiments

4.1 Datasets

We evaluate our method on four benchmark datasets covering diverse types of spurious correlations, following the experimental protocols established in recent robust learning literature (Liu et al., 2021; Qiu et al., 2023; Zhou and Zhu, 2025). The benchmarks include one toxicity detection task (CivilComments (Koh et al., 2021)), one natural language inference task (MultiNLI (Williams et al., 2018)), and two sentiment classification tasks with stylistic or conceptual confounds (Yelp-Author-Style and Beer-Concept-Occur (Zhou et al., 2024)). Detailed statistics, group configurations, and spurious correlation descriptions for each dataset are provided in Appendix C.1.



(a) Plain ERM exhibits entangled representations where style confounds sentiment.



(b) PRIFT produces a more clearly structured representation space while preserving some style variation within each class.

Figure 3: T-SNE visualization of latent space geometry on the Yelp-Author-Style dataset.

4.2 Baselines

We compare PRIFT against five baselines that do not require group labels: ERM (Liu et al., 2021), JTT (Liu et al., 2021), AFR (Qiu et al., 2023), CCR (Zhou and Zhu, 2025), and zero-shot-prompt. The zero-shot-prompt baseline uses the same E5 encoder as PRIFT, but directly performs classification with prompt-based representations without applying the PRIFT centered projection debiasing module. JTT trains a second model by upweighting samples misclassified by an initial ERM model. AFR retrains only the final layer of a fixed encoder by upweighting samples with high loss or

Methods	CivilComments			MultiNLI			Yelp-Author-Style			Beer-Concept-Occur		
	Mean	WGA	Gap	Mean	WGA	Gap	Mean	WGA	Gap	Mean	WGA	Gap
zero-shot-prompt	91.70	55.60	36.10	81.50	66.10	15.40	91.29	83.33	7.96	94.70	88.50	6.20
ERM	**92.60**	57.40	35.20	82.40	67.90	14.50	92.00	85.00	7.00	95.38	90.00	5.38
JTT	91.10	69.30	21.80	78.60	72.60	6.00	92.10	85.83	6.27	95.54	90.00	5.54
AFR	89.80	68.70	21.10	81.40	73.40	8.00	92.05	85.42	6.63	95.06	92.00	3.06
CCR	90.00	70.67	19.33	80.72	75.17	5.55	92.60	88.74	3.86	95.30	**93.04**	**2.26**
PRIFT	91.29	**74.12**	**17.17**	**84.11**	**79.63**	**4.48**	**93.16**	**91.15**	**2.01**	**95.79**	93.00	2.79

Table 1: Performance comparison with baselines that do not use instance-level group annotations. WGA, Mean Accuracy, and Accuracy Gap are reported across four datasets. The best score across all methods is **highlighted**.

low confidence. CCR employs causal inference and counterfactual reasoning to disentangle causal features from spurious ones. Additionally, we compare performance against two baselines that rely on group labels: Group-DRO (Sagawa et al., 2019), which minimizes the worst-group loss via explicit dynamic reweighting, and DFR (Kirichenko et al., 2023), which retrains the classifier head on a small dataset with group labels.

Methods	CivilComments			MultiNLI		
	Mean	WGA	Gap	Mean	WGA	Gap
Group-DRO	88.90	69.90	19.00	81.40	77.70	**3.70**
DFR	87.20	70.10	**17.10**	82.10	74.70	7.40
CCR	90.00	70.70	19.30	80.70	75.20	5.50
PRIFT	**91.29**	**74.12**	17.17	**84.11**	**79.63**	4.48

Table 2: Results comparing PRIFT with baselines with group labels across four datasets (CivilComments and MultiNLI).

4.3 Main results

Table 1 presents a comprehensive comparison against label-free baselines across four benchmarks with varying spurious correlations. PRIFT consistently achieves superior robustness; notably on the CivilComments dataset, it improves WGA by over 20% compared to ERM, while narrowing the accuracy gap from 35.20% to 13.17%. The exceptionally low average Equal Opportunity Gap (0.0062) and Equalized Odds Gap (0.0098) across all datasets confirm that these gains stem from genuine bias elimination rather than fortuitous variance reduction. Furthermore, Table 2 shows that PRIFT rivals "oracle" methods trained with full group supervision. Despite lacking sample-level group annotations, PRIFT attains a WGA of 78.12% on CivilComments (vs. 69.90% for Group-DRO), effectively bridging the performance gap between unsupervised and supervised robust learning .

The superior performance of PRIFT stems from the stability and precision of the centered projection mechanism, defined as $\hat{h} = h_x - (\alpha - \bar{\alpha}_s)v_s$. By normalizing the projection along the confounder direction v_s to the group mean $\bar{\alpha}_s$, this operator selectively suppresses the sample level deviations responsible for spurious correlations while preserving semantic information encoded in orthogonal directions. Moreover, the framework exhibits robustness and transferability due to the "plug-and-play" nature of prompt prototypes. Across all tasks, adapting PRIFT required only defining task-specific prompts for labels and confounders, without architectural modifications or extensive hyperparameter tuning. The consistent fairness metrics indicate that once appropriate geometric anchors are defined via prompts, the debiasing mechanism generalizes across diverse demographic and stylistic groups.

4.4 Investigating the impact of debiasing

To provide a qualitative view of how PRIFT reshapes the learned representation space, we visualize the Yelp-Author-Style test representations using t-SNE in Figure 3 as a qualitative illustration of representation reorganization. As shown in Figure 3, the representations learned by PRIFT are more clearly organized by sentiment, while style-related variation appears less visually prominent than in the ERM. Distinct style-specific subclusters may still persist within broader sentiment re-

Method	Debias	DRO	CivilComments		MultiNLI		Yelp-Author-Style		Beer-Concept-Occur	
			Mean	WGA	Mean	WGA	Mean	WGA	Mean	WGA
(a) Plain ERM	✗	✗	**92.60**	57.40	82.40	67.90	91.50	81.23	95.38	90.00
(b) Debias only	✓	✗	92.01	65.76	83.00	73.77	92.10	87.80	95.20	82.13
(c) DRO only	✗	✓	92.14	63.42	83.17	71.42	92.31	85.55	95.61	70.03
(f) Full model (ours)	✓	✓	91.29	**74.12**	**84.11**	**79.63**	**93.16**	**91.15**	**95.79**	**93.00**

Table 3: Ablation study analyzing the contributions of individual components.

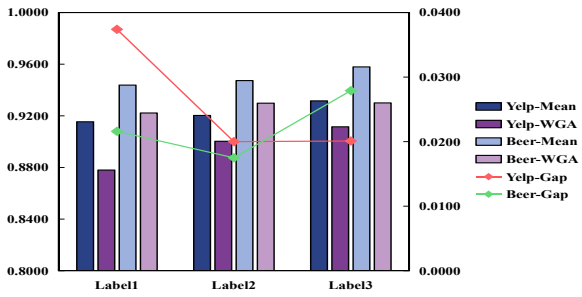
gions, suggesting that PRIFT does not completely erase style information. This observation is consistent with the design of centered projection, as the method suppresses instance-specific confounder deviations while preserving information in orthogonal directions rather than collapsing all confounder-related structure.

4.5 Ablation study

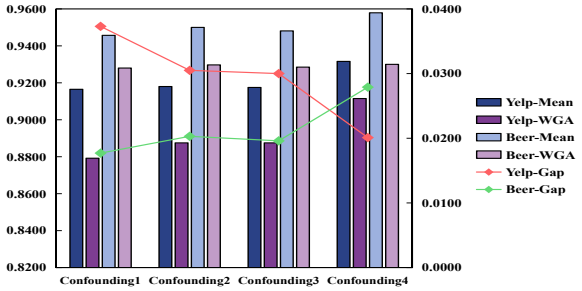
Table 3 summarizes the performance of these variants. Comparing (b) with (a) reveals that the centered projection mechanism independently drives robustness improvements, validating our hypothesis that model brittleness stems largely from latent geometric defects rather than solely optimization objectives. While DRO-only (c) outperforms ERM, it faces performance bottlenecks when applied directly to entangled representations where minority samples are geometrically indistinguishable from noise. In contrast, the Full Model achieves nonlinear performance gains that exceed the sum of individual components. We attribute this to a "purification-then-optimization" synergy: the centered projection acts as a feature purification step that removes dominant spurious deviations, thereby disentangling the latent space and enabling the robust optimization strategy to pinpoint and prioritize legitimate minority samples.

4.6 Prompt semantic sensitivity analysis

We analyze the sensitivity of label prompts using three templates summarized in Figure 4a: Label 1 ("The following is a [*dataset*] review that conveys a [*label*]."), Label 2 (which uses only the class label, e.g., "negative"), and Label 3 ("a [*dataset*] review expressing [*label*]."). Results indicate a clear performance hierarchy where Label 3 outperforms the others. While Label 1 is grammatically explicit, the introductory phrase ("The following is...") likely acts as semantic noise that dilutes



(a) Impact of different label prompts on accuracy.



(b) Impact of different confounding prompts on accuracy.

Figure 4: Sensitivity analysis of prompt semantic templates.

the discriminative signal in the label prototype v_l. Conversely, the superiority of Label 3 over the single-keyword Label 2 suggests that while keywords capture core sentiment, enclosing them in a concise, contextualized phrase helps the encoder better align the label prototype with sentence level review representations.

We further investigate confounding prompts using four variations shown in Figure 4b: Confounding 1 ("The following is a review written in [*confounding*] style."), Confounding 2 (which uses the confounding name directly, e.g., "Hemingway"), Confounding 3 (which uses descriptive features, e.g., "Short, direct, and concise sentences"), and Confounding 4 ("A text written in the style of [*style*]."). Consistent with label prompts, the concise template Confounding 4 achieves optimal per-

formance, reinforcing that minimizing semantic noise enhances prototype quality. A crucial finding emerges from the comparison between Confounding 2 and Confounding 3, which achieve nearly identical performance. This demonstrates that PRIFT does not rely on specific entity names; instead, it mitigates spurious correlations even when the specific confounder identity is unknown, provided that the confounding linguistic features can be described naturally.

5 Conclusion

In this paper, we propose PRIFT, a framework that mitigates spurious correlations by directly manipulating the geometry of the latent space through natural language prompts. Unlike approaches that relying on expensive group annotations or unstable adversarial training, PRIFT offers an interpretable, "plug-and-play" debiasing mechanism. Central to our approach is the centered projection operator, which purifies representations by suppressing deviations specific to confounders while preserving essential semantic structures. Additionally, the framework unifies varying levels of supervision, ranging from full group labels to unsupervised probes. For future work, we plan to explore automated methods for spurious concept discovery that identify unknown confounders without prompts defined by users, and to extend our geometric intervention strategies to multimodal tasks.

Limitations

Despite the robust performance of PRIFT, we acknowledge limitations related to its reliance on prior knowledge and the geometric assumptions of the framework. Specifically, optimal model performance relies on the ability to clearly describe spurious features using natural language prompts. Consequently, although the framework supports unsupervised inference mechanisms, the precision of geometric anchors cannot be optimized when confounders are unknown or cannot be described in natural language. In addition, PRIFT functions inherently as a first-order geometric intervention. The proposed centered projection operator assumes that the component of spurious correlations exploitable by classifiers primarily manifests as linear directional deviations. This design choice may limit the method's ability to address highly nonlinear entanglement issues. This is a common, unresolved problem when addressing spurious correlations. Furthermore, despite its high computational efficiency, this geometric intervention method is currently only applicable to discriminative tasks. While the proposed mechanism is not restricted to particular task types, the feasibility of extending it to generative tasks is a question for future research.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant nos. 2023YFB3308601, 2022YFB3104700), the National Natural Science Foundation (Grant nos. 62402395,62376198), Chengdu "Open bidding for selecting the best candidates" Science and Technology Project (Grant no. 2023-JB00-00020-GX), the Science and Technology Program of Sichuan Province (Grant no. 2023YFS0424), the Science and Technology Service Network Initiative (Grant no. KFJ-STS-QYZD-2021-21-001), and the Talents by Sichuan provincial Party Committee Organization Department, and Chengdu - Chinese Academy of Sciences Science and Technology Cooperation Fund Project (Major Scientific and Technological Innovation Projects)

References

John Aldrich. 1995. Correlations genuine and spurious in pearson and yule. *Statistical Science*, 10(4):364–376.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*.

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 of The 2019 World Wide Web Conference*, pages 491–500. ACM.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. 2021. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR.

Fatma Elsafoury, Stamos Katsigiannis, and Naeem Ramzan. 2023. On bias and fairness in NLP:

how to have a fairer text classification? *CoRR*, abs/2305.12829.

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

Karan Goel, Albert Gu, Sharon Li, and Christopher Ré. 2021. Model patching: Closing the subgroup performance gap with data augmentation. In *9th International Conference on Learning Representations*. OpenReview.net.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. 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, NAACL-HLT*, pages 107–112. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations*. OpenReview.net.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations,*. OpenReview.net.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, and 4 others. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5637–5664. PMLR.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via adapterfusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2730–2743.

Evan Zheran 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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6781–6792. PMLR.

Thien Hang Nguyen, Hongyang R. Zhang, and Huy L. Nguyen. 2023. Improved group robustness via classifier retraining on independent splits. *Trans. Mach. Learn. Res.*, 2023.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28448–28467. PMLR.

Aishik Rakshit, Smriti Singh, Shuvam Keshari, Arijit Ghosh Chowdhury, Vinija Jain, and Aman Chadha. 2025. From prejudice to parity: A new approach to debiasing large language model word embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6718–6747. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics.

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. *CoRR*, abs/1911.08731.

Yuanchao Wang, Zhao-Rong Lai, and Tianqi Zhong. 2025. Out-of-distribution generalization for total variation based invariant risk minimization. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

1112–1122. Association for Computational Linguistics.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2586–2614. Association for Computational Linguistics.

Yuqing Zhou and Ziwei Zhu. 2025. Fighting spurious correlations in text classification via a causal learning perspective. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4264–4274. Association for Computational Linguistics.

A Code

Our implementations for all experiments are available at `https://github.com/LuckyBot17/PRIFT`.

B Notation & Problem Setup

Table 4 summarizes the main notation used in the paper.

Symbol	Meaning
x	Input text instance
$l \in \mathcal{L}$	Label prompt describing a target class
$s \in \mathcal{S}$	Confounding prompt describing a nuisance attribute (e.g., style)
$g = (l, s) \in \mathcal{G}$	Group identity defined by label–confounder pair
$\mathcal{G} = \mathcal{L} \times \mathcal{S}$	Set of all groups
$f_\theta(\cdot)$	Text encoder applied to input texts
$f_{\text{enc}}(\cdot)$	Same encoder applied to prompts
$h_x = f_\theta(x)$	Raw text representation in latent space
$v_l = f_{\text{enc}}(l)$	Label prototype embedding
$v_s = f_{\text{enc}}(s)$	Confounder prototype embedding
$\alpha = \langle h_x, v_s \rangle$	Confounding intensity (projection onto v_s)
$\bar{\alpha}_s$	Mean confounding intensity for confounder s
W	Linear classifier parameters on top of debiased representations
$\hat{y}$	Model prediction (distribution over labels or groups, depending on head)
$\ell(\hat{y}, l)$	Training loss

Table 4: Notation used in the paper.

C Implementation

C.1 Datasets

We carefully selected these datasets to cover various types of spurious associations, ranging from demographic biases and syntactic artifacts to stylistic confounds. Table 5 describes the statistical characteristics and specific grouping details of each dataset.

CivilComments. CivilComments (Borkan et al., 2019) is a binary text classification task, where the task is to determine whether a comment is toxic. We adopt the version provided by the WILDS benchmark (Koh et al., 2021), which explicitly annotates mentions of demographic attributes such as gender and race. Spurious correlations arise from the co-occurrence of identity-related terms and toxicity labels in the training distribution: comments referencing specific demographic groups are disproportionately labeled as toxic, forming a "category-word shortcut". This dataset is used to evaluate whether PRIFT can reduce reliance on socially sensitive confounders and maintain performance stability on minority group samples where shortcut patterns do not hold.

MultiNLI. The Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018) dataset involves predicting the logical relationship between sentence pairs as entailment, contradiction, or neutral. A well documented spurious pattern exists between negation words and the contradiction label: models trained on standard splits often predict contradiction based solely on the presence of negation markers rather than semantic reasoning. Groups are formed by combining each of the three labels with negation presence or absence, resulting in six groups. This dataset tests whether PRIFT can prevent models from using shortcuts based on synonyms to correlate surface lexical features with inference labels.

Yelp-Author-Style. Yelp-Author-Style (Zhou et al., 2024) is a semi-synthetic sentiment classification dataset, which is derived from Yelp reviews by rewriting texts in distinct literary styles of Hemingway and Shakespeare while preserving sentiment polarity. We simplify the task to binary classification by selecting reviews with ratings of 2 (negative) and 4 (positive). The spurious correlation is introduced artificially: lower ratings are predominantly associated with Hemingway's concise style, while higher ratings align with Shakespeare's elaborate prose. Groups correspond to the four combinations of sentiment and authorial style. This dataset enables controlled analysis of how geometric debiasing handles author-style confounders that are orthogonal to sentiment semantics.

Beer-Concept-Occur. Beer-Concept-Occur (Zhou et al., 2024) is constructed

Dataset	GID	Label	Group (spurious, class)	#Train
CivilComments	0	0	(no identities, non-toxic)	148,486
	1	0	(has identities, non-toxic)	90,337
	2	1	(no identities, toxic)	12,731
	3	1	(has identities, toxic)	17,784
MultiNLI	0	0	(no negations, contradiction)	57,498
	1	0	(has negations, contradiction)	11,158
	2	1	(no negations, entailment)	67,376
	3	1	(has negations, entailment)	1,521
	4	2	(no negations, neutral)	66,630
	5	2	(has negations, neutral)	1,992
Yelp-Author-Style	0	0	(Hemingway, negative)	684
	1	0	(Shakespeare, negative)	214
	2	1	(Hemingway, positive)	252
	3	1	(Shakespeare, positive)	660
Beer-Concept-Style	0	0	(Appearance, rating 0.6)	477
	1	0	(Aroma, rating 0.6)	261
	2	1	(Appearance, rating 1.0)	65
	3	1	(Aroma, rating 1.0)	432

Table 5: Statistics and group configurations of the four benchmark datasets.

from beer reviews; the binary classification task predicts palate ratings based on review text. Each sample contains a focused review of the palate alongside a segment that confounds the review by describing either the beer's aroma or appearance. The spurious correlation emerges from the association between comments related to appearance and low palate ratings, in contrast to the link between comments related to aroma and high palate ratings. This dataset is used to test whether PRIFT can ignore specific domain concepts and focus on causal sentiment indicators.

Hyperparameter	Value
Pretrained encoder	e5-base-v2
Encoder batch size	64
Learning rate (linear head)	5×10^{-3}
Weight decay (linear head)	1×10^{-4}
Epochs (linear head)	25
Random seeds	5, 17, 20, 42, 66

Table 6: Hyperparameter settings.

C.2 Hardware devices and parameter configuration

All experiments are conducted on a single workstation equipped with one NVIDIA RTX 4090 GPU and 16 vCPUs IntelR XeonR Gold 6430 processor. We use the same pretrained encoder and optimization hyperparameters across all four datasets. In particular, we adopt e5-base-v2 as the text encoder, and use a batch size of 64 for encoding the inputs. On top of the debiased sentence representations, we train a linear classification head with a fixed configuration: the head is optimized using AdamW with a learning rate of 5×10^{-3}, a weight decay of 1×10^{-4}, and is trained for 25 epochs. To assess robustness and ensure reproducibility, we repeat all experiments with five different random seeds, $\{5, 17, 20, 42, 66\}$, which are used consistently for data shuffling and parameter initialization. Unless otherwise specified, the results reported in this paper are the average values across these seeds.

Importantly, we employ a single set of hyperparameters for all four benchmarks, without any tuning specific to each dataset. This design choice reflects the fact that our framework does not rely on hyperparameter search conducted for each dataset to achieve robust performance and remains stable under a unified configuration. A summary of the key hyperparameters is provided in Table 6.

C.3 Evaluation metric

Assessing model robustness to spurious correlations requires metrics that capture both aggregate performance and the quality of predictions across distribution shifts and underrepresented subgroups. We adopt a comprehensive evaluation framework that combines group-level accuracy measures with fairness-oriented robustness indicators, providing a multifaceted view of model behavior under spurious correlation.

Mean Accuracy. The overall accuracy across the entire test set, computed as the fraction of correctly classified samples. This metric reflects standard predictive performance but can mask severe disparities across groups when spurious correlations are present.

Worst-Group Accuracy (WGA). The minimum accuracy achieved over all groups, formally defined as the minimum group accuracy,

$$\text{WGA} = \min_{g \in G} \text{Acc}(g) \tag{14}$$

where $g = (l, s)$ denotes a group defined by l and s. WGA directly measures the extent to which a model maintains performance on the most challenging subgroup, serving as the primary robustness indicator. High WGA affected that the model does not disproportionately fail on minority groups induced by spurious correlations.

Gap. The difference between mean accuracy and WGA,

$$\text{Gap} = \text{Mean Accuracy} - \text{WGA} \tag{15}$$

This metric quantifies the disparity between average and worst-case performance. Smaller gaps indicate more uniform prediction quality across groups, reflecting reduced reliance on spurious features that advantage certain subgroups over others.

Equal Opportunity Gap (EO Gap). For binary classification with positive label $y = 1$, equal opportunity requires that the true positive rate be equal across groups: $TPR(g_1) = TPR(g_2)$ for all groups $g_1, g_2 \in G$. The EO Gap measures the maximum disparity in true positive rates,

$$\text{EO Gap} = \max_{g_1, g_2 \in G} |\text{TPR}(g_1) - \text{TPR}(g_2)| \tag{16}$$

In the context of spurious correlation mitigation, this metric serves as a robustness complement rather than solely a fairness constraint: it detects whether the model's ability to correctly identify positive instances is artificially inflated or deflated by confounder-label associations. A lower EO Gap indicates that positive class predictions are robust to confounder variations.

Equalized Odds Gap (EOdds Gap). Equalized odds extends equal opportunity by requiring both true positive rates and false positive rates to be equalized across groups. The Equalized Odds Gap is defined as,

$$\begin{aligned}\text{EOdds Gap} = \max_{g_1, g_2 \in \mathcal{G}} \Big(&|\text{TPR}(g_1) - \text{TPR}(g_2)| \\ &+ |\text{FPR}(g_1) - \text{FPR}(g_2)| \Big)\end{aligned} \tag{17}$$

where FPR denotes the false positive rate. This metric captures whether the model's confusion patterns vary systematically with confounders. In our robustness evaluation framework, we interpret EOdds Gap as a diagnostic for spurious feature dependence: models that exploit shortcuts typically exhibit large disparities in both their sensitivity to true positives and their propensity to generate false alarms across confounder-defined groups. Consequently, we treat EO Gap and EOdds Gap as supplementary robustness indicators that provide orthogonal evidence of spurious correlation mitigation, rather than as fairness-only criteria.

D Comparison of Confounder Assignment Strategies

Table 7 compares the three confounder assignment options in Eq. 7. Across both datasets, all PRIFT variants substantially outperformed ERM. This indicates that centered projection debiasing is beneficial even when confounder information is incomplete. Notably, the probe-based setting with 100% confounder labels achieves a WGA nearly identical to using true s, suggesting that accurate confounder assignment is sufficient for PRIFT to approach its robustness upper bound. Crucially, our prototype inference mechanism requires zero confounder labels yet remains competitive. On CivilComments, it achieves 74.12%, outperforming the probe-based setting with only 10% confounder labels (70.37%) by 3.75%. This suggests that prototype-based assignments are more reliable than lightly supervised probes under limited supervision. On Yelp-Author-Style, prototype inference is comparable to the 10% probe and remains close to the true s setting. Overall, PRIFT provides a practical plug-and-play solution: without additional confounder annotation, prototype inference closes a large fraction of the robustness gap between ERM and fully super-

Setting	CivilComments WGA	Yelp WGA
PRIFT (true s)	78.19	92.40
PRIFT (prototype inference; label-free)	74.12	91.15
PRIFT (probe-based; 10% s labels)	70.37	91.00
PRIFT (probe-based; 100% s labels)	77.66	92.10
ERM (no debiasing)	57.40	85.00

Table 7: WGA under different confounder assignment strategies (Eq. 7) on CivilComments and Yelp-Author-Style. **PRIFT (true s)** uses oracle confounder labels to assign s; **PRIFT (prototype inference)** infers s by nearest confounder prototype (label-free); **PRIFT (probe-based)** predicts s with a learned probe trained using either 10% or 100% confounder-labeled instances; **ERM** denotes standard empirical risk minimization without debiasing.

vised variants.

E Verification of semantic alignment

A core principle of the PRIFT framework is that the classifier is not learned from zero as an abstract decision boundary; instead, it is explicitly guided by the semantic embeddings of the label prompts. To verify this semantic dependency and rule out the possibility that the model is merely memorizing class indices while ignoring the prompt content, we conducted experiments that reversed the correspondence between label prompts and ground truth classes.

Specifically, we swapped the prompts for the positive and negative classes (e.g., we initialized the weights for the positive class using the "negative" label prototype and vice versa). Table 8 presents the results for the sensitivity analysis on the Yelp-Author-Style and Beer-Concept-Occur datasets. Replacing the label prompts with their opposites substantially decreases accuracy. In the true order setting, the model achieves a high WGA of 91.15% on Yelp-Author-Style and 93.00% on Beer-Concept-Occur. However, under the reverse order setting, the WGA drops to 13.33% and 19.91%, respectively. These accuracy scores are well below the 50% threshold of random guessing for binary classification. This suggests that the model is not merely confused; it is confidently making predictions that align with the true order prompts. This future suggests that the decision boundaries in

PromptTemplate	Yelp-Author-Style		Beer-Concept -Occur	
	Mean	WGA	Mean	WGA
Reverse order	34.25	13.33	35.11	19.91
True order	93.16	91.15	95.79	93.00

Table 8: Performance under different prompt orders.

PRIFT are closely aligned with the latent semantic directions defined by the prototypes. This supports that the prompt prototypes act as effective geometric anchors, and the model is robust because it can understand and use these semantic directions.

F LLM-in-the-loop confounder discovery

To further test whether PRIFT can operate without manually written confounder prompts, we conduct an LLM-in-the-loop discovery experiment. In this setting, we use frozen large language models to read the training data and generate confounder prompts automatically. These prompts are then encoded in the same way as our designed prototypes and are used by PRIFT. This setting evaluates whether the proposed framework can maintain robustness when the discovery of spurious cues is delegated to an external text-only generator.

Table 9 reports results on CivilComments and Yelp-Author-Style. Among the frozen LLMs, GPT-4o yields the strongest performance, achieving 87.38% WGA on CivilComments and 89.98% WGA on Yelp-Author-Style, which suggests that stronger instruction-following models can produce more informative confounder descriptions. At the same time, PRIFT with manually specified confounder prompts remains the best overall method, reaching 88.12% and 91.15% WGA on the two datasets, respectively. This indicates that the automatically discovered prompts capture a substantial portion of the relevant spurious factors, although they still fall short of carefully designed prompts.

G Algorithm for centered projection debiasing

Algorithm 1 outlines the core geometric debiasing procedure of the PRIFT framework. The algorithm operates in two main phases. In the estimation phase, it iterates through the training dataset to compute the confounder-wise mean confounding intensity $\bar{\alpha}_s$ for each confounder prototype v_s.

Methods	CivilComments		
	Mean	WGA	Gap
LLaMA3-8B-instruct	90.98	86.79	4.19
flan-t5-xl	90.76	86.11	4.65
GPT-4o	91.14	87.38	3.76
PRIFT	**91.29**	**88.12**	**3.17**
Methods	**Yelp-Author-Style**		
	Mean	**WGA**	**Gap**
LLaMA3-8B-instruct	92.64	89.17	3.47
flan-t5-xl	92.55	88.34	4.21
GPT-4o	92.70	89.98	2.72
PRIFT	**93.16**	**91.15**	**2.01**

Table 9: Results of LLM-in-the-loop confounder discovery. Frozen LLMs generate confounder prompts directly from training texts without human intervention, and the discovered prompts are then used within PRIFT.

This mean value serves as a confounder-level reference intensity for v_s, capturing the average alignment of instances with confounder s in the training data. In the projection phase, the algorithm applies the centered projection operator to each input instance. By calculating the instance-specific intensity $\alpha = \langle h_x, v_s \rangle$ and subtracting $\lambda(\alpha - \bar{\alpha}_s)v_s$ from h_x, it produces debiased representations $\hat{h}$. This transformation preserves the components of h_x orthogonal to v_s while reducing deviations specific to each instance along the confounder direction. These deviations are often associated with spurious shortcuts.

H The Use of Large Language Models

We used a large language model only to refine the written paragraphs, making them more fluent and readable. No other aspects of the work used large language models beyond this text refinement.

Algorithm 1 Centered Projection Debiasing in Latent Space

Require: Encoder f_θ, confounder prototypes $\{v_s\}_{s\in\mathcal{S}}$, training set $\mathcal{D} = \{(x_i, l_i, s_i)\}_{i=1}^{N}$, debiasing strength λ

Ensure: Debiased representations $\{\hat{h}\}$ for all x

Normalize v_s:
for each $s \in \mathcal{S}$ **do**
  $v_s \leftarrow v_s / \|v_s\|_2$
end for
Initialize accumulators for confounder-wise mean intensity $\bar{\alpha}_s$:
for each $s \in \mathcal{S}$ **do**
  $M_s \leftarrow 0$ ▷ accumulator
  $C_s \leftarrow 0$ ▷ counter
end for
Estimate confounder-wise mean intensity $\bar{\alpha}_s$:
for each mini-batch $B \subset \mathcal{D}$ **do**
  for each $(x, l, s) \in B$ **do**
    $h_x \leftarrow f_\theta(x)$
    Assign confounder index $s \leftarrow \gamma(x)$ using the chosen strategy ▷ optional ℓ_2-normalization
    $\alpha \leftarrow \langle h_x, v_s \rangle$ ▷ confounding intensity
    $M_s \leftarrow M_s + \alpha$, $C_s \leftarrow C_s + 1$
  end for
end for
for each $s \in \mathcal{S}$ **do**
  $\bar{\alpha}_s \leftarrow M_s / C_s$ ▷ confounder-wise mean intensity
end for
Apply centered projection to obtain debiased features:
for each batch $B \subset \mathcal{D}$ **do**
  for each $(x, l, s) \in B$ **do**
    $h_x \leftarrow f_\theta(x)$
    $h_x \leftarrow h_x / \|h_x\|_2$
    $\alpha \leftarrow \langle h_x, v_s \rangle$
    $\delta \leftarrow \alpha - \bar{\alpha}_s$ ▷ deviation from confounder-wise mean
    $\hat{h} \leftarrow h_x - \lambda\, \delta\, v_s$ ▷ centered projection
    $\hat{h} \leftarrow \hat{h} / \|\hat{h}\|_2$ ▷ optional re-normalization
    Store $\hat{h}$ for subsequent classifier training
  end for
end for
return $\{\hat{h}\}$