EDGE: Enhanced Debiased Gradient Extraction for Robust Fine-tuning

Jinglong Li[1], **Kun Zhang**[1,*], **Chenyu Zou**[2], **Wei Shi**[1], **Xin Li**[3,4], **Si Wei**[4]
[1] School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[2] College of Art and Design, Nanjing Tech University, Nanjing 211816, China
[3] School of Information Science and Technology,
University of Science and Technology of China, Hefei 230026, China
[4] Artificial Intelligence Research Institute, iFLYTEK Co., Ltd., Hefei 230088, China
{lijinglong.edu, zhang1028kun, zouchenyu}@gmail.com,
singeraw@mail.hfut.edu.cn, leexin@ustc.edu.cn, siwei@iflytek.com

Abstract

Recent advances in large-scale pre-training have substantially enhanced the robustness and generalization capabilities of foundation models (e.g., Qwen3 and Llama-4). However, when fine-tuning them on downstream tasks, these models often latch onto dataset-specific biases, learning spurious correlations tied to *easy-to-learn* but non-robust features. This undermines their performance under distribution shifts, despite strong in-distribution (ID) accuracy. Existing fine-tuning methods, including full-parameter and parameter-efficient techniques, primarily optimize for ID performance and largely overlook out-of-distribution (OOD) robustness. Meanwhile, debiasing has been explored in full fine-tuning, while debiasing strategies on Parameter-Efficient Fine-Tuning (PEFT) remain underexplored. To this end, in this paper, we propose *Enhanced Debiased Gradient Extraction* (EDGE), a lightweight gradient projection-based method that explicitly suppresses bias-amplifying updates during fine-tuning process. EDGE is a model-agnostic, and plug-and-play debiasing method that operates without relying on predefined bias types or labels. It seamlessly integrates with both full and parameter-efficient fine-tuning, and generalizes across NLP and vision tasks. Experiments on synthetic and real-world benchmarks demonstrate that EDGE effectively reduces bias and consistently improves OOD generalization, offering a unified and practical framework for robust adaptation under dataset bias.

Keywords: Natural language inference , Spurious correlations , Robust Fine-tuning , Debiased reasoning

1 Introduction

The emergence of large-scale foundation models—such as Qwen3 (Team, 2025), Llama-4 (AI@Meta, 2025), and Gemma 3 (Team et al., 2025)—has significantly advanced the frontier of general artificial intelligence. These models exhibit remarkable generalization and adaptability across a wide range of tasks, particularly in natural language processing and complex reasoning. Their success stems from two complementary strengths: (1) strong generalization capabilities acquired through pre-training; and (2) high adaptability to specific tasks via downstream fine-tuning.

As the mainstream of adapting pre-trained models to downstream tasks, fine-tuning strategies still suffer from critical problems. Most existing fine-tuning strategies focus heavily on optimizing in-distribution (ID) performance, often neglecting their impact on out-of-distribution (OOD) generalization (Liu et al., 2024; Zhang et al., 2025b). In practice, fine-tuning attempts to use as little downstream data as possible to achieve efficient capability transformation. Thus, domain-specific artifacts and spurious correlations contained in the fine-tuning data may be abused by models since they are *easy to learn* (Wang et al., 2024). For full fine-tuning, despite its high computational costs, massive parameter updating will cause pre-trained models to overfit easy-to-learn spurious features, resulting in poor generalization under distribution shifts (Kumar et al., 2022). In contrast, parameter-efficient fine-tuning

* Corresponding Author

Proceedings of the 24th China National Conference on Computational Linguistics, pages 890-903, Jinan, China, August 11-14, 2025.

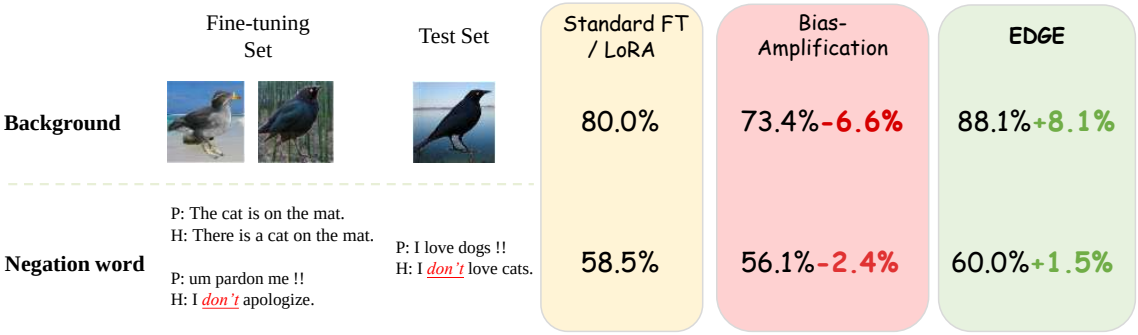

Figure 1: Performance comparison across vision and language domains under spurious attribute conditions. In the CV setting, models are evaluated on the Waterbirds dataset, where background correlates spuriously with class labels; In the NLI setting, models are fine-tuned on SNLI and evaluated on HANS.

(PEFT) provides a more scalable alternative by updating only a small subset of parameters. However, due to its constrained update space and limited representational flexibility, PEFT also struggles to effectively reorient model behavior away from spurious correlations learned during pre-training or introduced by biased fine-tuning data (Das et al., 2024). As a result, they may implicitly preserve or even reinforce existing biases, despite their efficiency advantages. As shown in Figure 1, models can achieve high ID accuracy by exploiting *easy-to-learn* cues—such as background textures in vision tasks or negation tokens in natural language inference (NLI)—instead of acquiring robust features. Such biased learning will collapse the model capability when dealing with distribution shifts. Recent efforts (Wortsman et al., 2022b; Zhu et al., 2023a; Wang et al., 2024; Tian et al., 2024) explore ways to mitigate bias introduced during fine-tuning or to preserve pre-trained knowledge for better generalization. However, these approaches predominantly rely on full-parameter fine-tuning, limiting their practicality in PEFT strategies and dealing with computationally and memory-intensive large language models. This raises an important question: "**How to realize unbiased fine-tuning of pre-trained models in full-parameter and parameter-efficient scenarios?**"

To this end, in this paper, we introduce *Enhanced Debiased Gradient Extraction* (EDGE), a simple yet effective fine-tuning strategy that explicitly removes bias gradients to improve the robustness under different fine-tuning scenarios. Specifically, at each tuning step, EDGE computes two gradient signals: (1) the fine-tuning gradient, which reflects both core and spurious task signals and (2) a bias-amplified gradient (Nam et al., 2020), obtained by fine-tuning on data designed to emphasize spurious correlations. We extract the basic gradient as the shared component between the bias-amplified and fine-tuning gradients, reflecting the model's generalizable knowledge. The remaining part of the bias-amplified gradient—orthogonal to the basic gradient—is defined as the bias gradient, which captures the direction of spurious features such as background cues. We then project the original fine-tuning gradient to remove its bias component, yielding the EDGE gradient. This adjusted direction focuses updates toward robust, task-relevant features —while suppressing updates aligned with spurious cues. This mechanism enables EDGE to mitigate shortcut learning while retaining essential task knowledge. Notably, EDGE is simple, model-agnostic, and seamlessly integrates with both standard fine-tuning and parameter-efficient pipelines, without requiring any predefined bias types or annotations. We validate EDGE through comprehensive experiments: (1) a controlled bias-injection protocol that allows systematic assessment of debiasing effectiveness, (2) real-world robustness benchmarks on NLP and CV under natural distribution shifts. In all cases, EDGE consistently improves robustness while maintaining core task performance. We publicly release our code to facilitate further research and practical adoption at: https://github.com/qingli-ql/EDGE.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 890-903, Jinan, China, August 11-14, 2025.

2 Related Work

2.1 Parameter-Efficient Fine-Tuning

PEFT allows pre-trained models to rapidly adapt to new tasks with minimal additional parameters. Prominent methods include BitFit (Zaken et al., 2021), Adapters (Houlsby et al., 2019), Prompt Tuning (Wang et al., 2023; Xiong et al., 2024), Prefix Tuning (Li and Liang, 2021), and Low-Rank Adaptation (LoRA) (Hu et al., 2021; Valipour et al., 2022; Zhang et al., 2023; Liu et al., 2024; Zhang et al., 2025a; Zhang et al., 2025b). Among these, LoRA stands out due to its computational efficiency, as it employs trainable low-rank matrices to approximate weight adjustments with reduced computational overhead.

While PEFT approaches have made significant progress, they often fail to account for an essential challenge: the potential biases introduced during fine-tuning. When models are adapted to imbalanced or spurious datasets, these biases can lead to a notable decline in performance.

2.2 Fine-tuning Pre-trained Models for Robustness

Fine-tuning pre-trained models often risks degrading their generalization capabilities(Zhang et al., 2025c). Kumar et al. (2022) introduced the Feature Distortion Theory, which explains the differences in feature space behavior between ID and OOD examples during standard full fine-tuning. They proposed the ”first linear probing, then fine-tune” strategy, which mitigates feature distortion and preserves generalization.

Subsequent methods for robust fine-tuning include: Selective Layer Fine-Tuning: Fine-tuning specific layers of the pre-trained model to limit unnecessary adjustments (Shen et al., 2021; Lee et al., 2022). Controlling Model Distance: Techniques that constrain the distance between the pre-trained and fine-tuned models to maintain learned representations (Xuhong et al., 2018; Gouk et al., 2020; Tian et al., 2023; Tian et al., 2024). Ensemble-based Approaches: Combining pre-trained and fine-tuned models to leverage their respective strengths (Wortsman et al., 2022a; Wortsman et al., 2022b; Wang et al., 2024). Guided Fine-Tuning: Using the pre-trained model as a guide during fine-tuning to align with the pre-trained generalization direction (Zhu et al., 2023a; Zhu et al., 2023b).

While these strategies address robustness in full fine-tuning scenarios, the domain of PEFT remains largely unexplored. Developing robust PEFT methods to maintain generalization while adapting to downstream tasks represents a critical area for future research.

2.3 Debiased Learning with Known Bias

Debiasing methods often rely on prior knowledge of biases to reduce model dependence on spurious correlations. Key strategies include: Data-level methods mitigate bias by modifying or augmenting training data (Schuster et al., 2021; Wu et al., 2022). Regularization-based methods introduce auxiliary objectives or components to penalize biased representations, often requiring multi-stage training or additional models (Utama et al., 2020; Du et al., 2021; Lyu et al., 2023). Reweighting and resampling approaches adjust sample importance to counteract bias (Sagawa et al., 2020; Jang and Wang, 2023), while adversarial training promotes invariance to biased features by discouraging reliance on spurious signals (Moyer et al., 2018; Kim et al., 2019; Lim et al., 2023). Causal inference techniques aim to disentangle bias from causal features using counterfactual reasoning (Niu et al., 2021; Zhang et al., 2024a; Zhang et al., 2024b). Furthermore, invariant learning (Arjovsky et al., 2019; Sagawa et al., 2019; Krueger et al., 2021; Yang et al., 2025; Liao et al., 2025) strives to trains models to maintain consistent performance across varying bias attributes.

Our Distinction. Existing methods mainly focus on improving in-distribution accuracy, often overlooking biases introduced during fine-tuning. In contrast, we propose EDGE, a lightweight debiasing strategy that corrects spurious correlations during adaptation with minimal additional parameters. EDGE naturally integrates with both full-parameter and parameter-efficient fine-tuning (e.g., LoRA), making it practical for real-world scenarios where training data are often biased. By suppressing shortcut learning, EDGE preserves the pre-trained model's robustness under distribution shifts.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 890-903, Jinan, China, August 11-14, 2025.

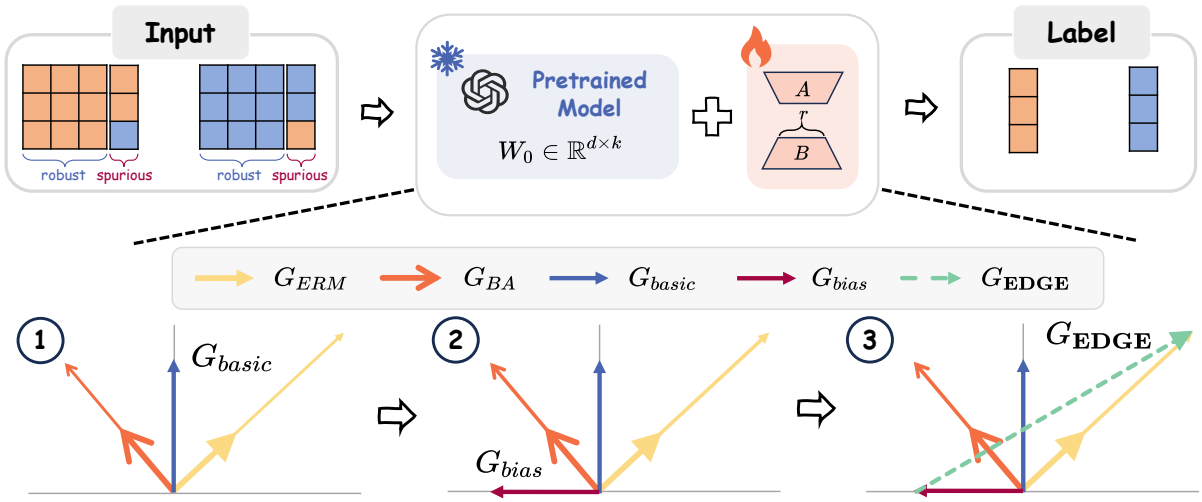

Figure 2: Overall framework of EDGE in a 2D vector-space perspective, exemplified under LoRA-based fine-tuning. ① Common Subspace Identification; ② Hyperplane Projection and Bias Gradient Identification; ③ Debiased Gradient Computation. "BA Gradient" denotes the Bias-Amplified Gradient.

3 Preliminary

3.1 LoRA-based PEFT

LoRA (Hu et al., 2021; Liu et al., 2024) stands out as one of the most representative methods within the PEFT framework, distinguished by its *computational efficiency* and its ability to approximate the performance of full-parameter fine-tuning. This technique leverages the product of two trainable low-rank matrices to approximate weight updates while minimizing computational costs.

For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA approximates the weight update $\Delta W \in \mathbb{R}^{d\times k}$ using a low-rank factorization, represented as BA. Here, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are the two low-rank matrices, with $r \ll \min(d, k)$. The resulting fine-tuned weight W' is then defined as:

$$W' = W_0 + \Delta W = W_0 + \underline{BA}. \tag{1}$$

Here, W_0 remains fixed during fine-tuning, while the parameters of A and B are updated. The matrix A is initialized using a uniform Kaiming distribution (He et al., 2015), and B is initialized to zero, ensuring that $\Delta W = \underline{BA}$ starts zero at the beginning of training.

3.2 Problem Definition: Debiased Fine-tuning

In standard supervised learning, models are trained to minimize the Cross Entropy (CE) loss under the Empirical Risk Minimization (ERM) framework:

$$L_{\text{ERM}}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{2}$$

where y is the ground-truth label and $\hat{y}$ is the predicted probability distribution over C classes. The corresponding gradient with respect to model parameters θ is given by

$$G_{\text{ERM}} = \frac{\partial L_{\text{ERM}}}{\partial \theta} = -\frac{1}{\hat{y}_c}\frac{\partial \hat{y}_c}{\partial \theta}, \tag{3}$$

where c denotes the correct class index.

However, when fine-tuning on biased datasets (Nam et al., 2020; Shah et al., 2020; Tiwari and Shenoy, 2023), ERM gradients often exploit spurious, easy-to-learn features that fail under distribution shifts. This motivates us to redefine fine-tuning from a gradient perspective: ***debiasing aims to remove bias-related components from the update direction***.

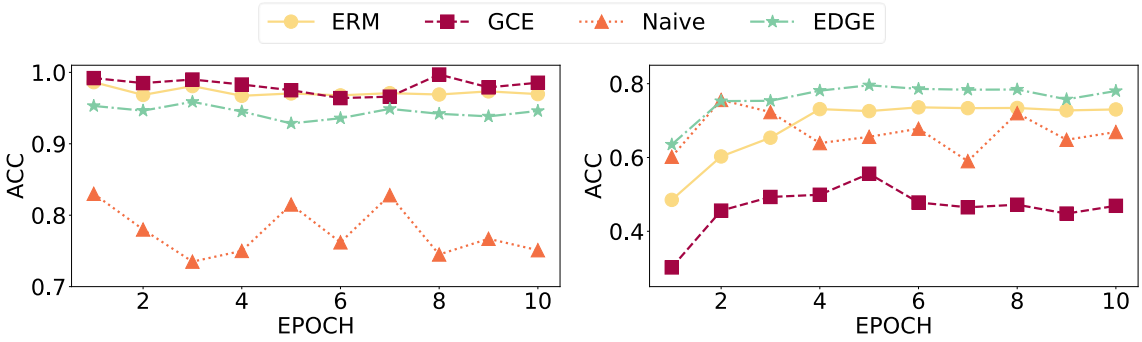

Figure 3: Accuracy Curves of RoBERTa-base on the Biased SNLI Dataset under Varying Gradient Update. (Left) Average accuracy for class c_k instances that contain the spurious token t_s. (Right) Average accuracy for instances not belonging to class c_k but still containing the spurious token t_s.

Specifically, we seek a debiased gradient G_{target} by subtracting the projection of G_{ERM} onto the bias-dominant subspace:

$$G_{\text{target}} = G_{\text{ERM}} - \Pi_{\text{bias}}(G_{\text{ERM}}), \tag{4}$$

where $\Pi_{\text{bias}}(\cdot)$ denotes the projection operator. Such refinement encourages learning robust, task-relevant features rather than superficial correlations.

4 Technical Details of EDGE

In this paper, we propose a debiasing framework, *Enhanced Debiased Gradient Extraction* (EDGE), as shown in Figure 2. Inspired by the success of bias amplification and mitigation techniques (Nam et al., 2020; Ahn et al., 2022), we begin with a comprehensive analysis of gradient update dynamics across various fine-tuning strategies. This analysis highlights how fine-tuning can inadvertently amplify spurious correlations, providing insights for our method's design. Our goal is to identify and remove bias-related components at the gradient level. To achieve this, we use gradient projection to ensure the model focuses on task-relevant features, minimizing reliance on spurious correlations. The following sections will introduce each of these steps in detail.

4.1 Identifying and Decomposing the Bias-Driven Gradient

To identify bias-related gradient components within G_{ERM}, we follow prior works (Nam et al., 2020; Ahn et al., 2022), which suggest amplifying bias during training to expose its influence on optimization. Specifically, we use the Generalized Cross Entropy (GCE) loss (Zhang and Sabuncu, 2018) to enhance bias effects:

$$L_{\text{GCE}}(y, \hat{y}) = \frac{1 - (p_y)^q}{q}, \tag{5}$$

$$G_{GCE} = \frac{\partial L_{\text{GCE}}(y, \hat{y})}{\partial \hat{y}} = -(p_y)^q \frac{1}{\hat{y}_c} \frac{\partial \hat{y}_c}{\partial \theta} = (p_y)^q G_{\text{ERM}}. \tag{6}$$

where p_y is the predicted probability for the correct class, and q controls the extent of bias amplification. Unlike the standard CE loss, GCE scales the gradient by $(p_y)^q$, emphasizing easier examples and amplifying spurious correlations. This makes G_{GCE} a reliable indicator of the bias direction. Then, a naive approach to debiasing involves removing the direction of G_{GCE} from G_{ERM} via orthogonal projection:

$$G_{\text{naive}} = G_{\text{ERM}} - G_{\text{ERM}} \frac{G_{\text{GCE}}}{|G_{\text{GCE}}|}. \tag{7}$$

Dataset	Focus Category	Size
QNLI (Wang et al., 2018)	Out-of-distribution	5,266
MNLI-hard-m (Mahabadi et al., 2019)	Out-of-distribution	4,573
MNLI-hard-mm (Mahabadi et al., 2019)	Out-of-distribution	4,530
ST (Naik et al., 2018)	Stress Test (distraction & noise)	93,447
HANS (McCoy et al., 2019)	Syntactic Heuristic	30,000
IS-CS (Nie et al., 2019)	Inter-sentences Heuristics	654

Table 1: Evaluation Datasets for SNLI-Generalization Experiments

However, as shown in Figure 3, this method disrupts convergence and degrades performance, indicating that G_{GCE} is not purely bias-driven but also contains task-relevant components. To address this, we decompose G_{GCE} into two orthogonal components:

$$G_{\text{GCE}} = \underbrace{\alpha\, G_{\text{basic}}}_{\text{Core Feature Component}} + \underbrace{\beta\, G_{\text{bias}}}_{\text{Bias Component}}, \tag{8}$$

where G_{basic} captures essential robust features and G_{bias} encodes spurious correlations. The orthogonality assumption (Shah et al., 2020; Joshi et al., 2022) $\langle G_{\text{basic}}, G_{\text{bias}} \rangle = 0$ ensures debiasing removes bias without impeding task-relevant learning, allowing the model to focus on robust patterns. Building on this, we present our proposed debiased fine-tuning framework: EDGE.

4.2 The Debiasing Operation of EDGE

① *Common Subspace Identification:* Given that a unique hyperplane exists between any two linearly independent vectors, we designate this hyperplane as the basic feature space. In this space, the projection of the bias-amplified gradient G_{GCE} captures the core linguistic features. To extract the basic feature direction, we compute the angular bisector of G_{ERM} and G_{GCE}, which serves as a consensus direction:

$$G_{\text{basic}} = \frac{G_{\text{ERM}}}{|G_{\text{ERM}}|} + \frac{G_{\text{GCE}}}{|G_{\text{GCE}}|}. \tag{9}$$

② *Hyperplane Projection and Bias Gradient Identification:* We next isolate the bias component in G_{GCE} by projecting it onto the common feature space G_{basic}. This decomposition splits G_{GCE} into two orthogonal components: one along G_{basic} and the other capturing the residual bias. The component of G_{GCE} along G_{basic} is given by:

$$G_{\text{GCE}}^{\text{basic}} = \frac{\langle G_{\text{GCE}},\, G_{\text{basic}} \rangle}{\langle G_{\text{basic}},\, G_{\text{basic}} \rangle}\, G_{\text{basic}}. \tag{10}$$

The bias gradient is then obtained by subtracting this projection:

$$G_{\text{GCE}}^{\text{bias}} = G_{\text{GCE}} - G_{\text{GCE}}^{\text{basic}}. \tag{11}$$

This decomposition isolates the bias-related component from the core linguistic features, enabling targeted debiasing.

③ *Debiased Gradient Computation:* The debiased gradient is computed by removing the bias component from G_{ERM}, with α controlling the extent of debiasing:

$$G_{\textbf{EDGE}} = G_{\text{ERM}} - \alpha\, G_{\text{GCE}}^{\text{bias}}. \tag{12}$$

5 Experiments

In this section, we detail our experimental setup and evaluation protocol. We begin by introducing the datasets used across tasks, followed by descriptions of baseline methods and key hyperparameter settings. Finally, we present the main results along with in-depth analysis. All models are evaluated using accuracy % on various test sets, with bold and underline indicating the **best** and second-best results, respectively

Backbone	Methods	$p_{prev}, p_{str} = 0.25, 0.90$		$p_{prev}, p_{str} = 0.33, 0.50$	
		OA	WGA	OA	WGA
RoBERTa-base	ERM	0.812	0.735	0.832	0.786
	REW	0.814	0.748	<u>0.835</u>	0.791
	LM	0.801	0.740	0.812	0.785
	CORSAIR	0.810	0.732	0.827	0.780
	GA	<u>0.824</u>	<u>0.773</u>	0.830	**0.793**
	EDGE	**0.835**	**0.774**	**0.843**	<u>0.792</u>
GPT2	ERM	0.718	0.642	0.740	0.693
	REW	0.720	0.646	0.743	0.689
	LM	0.711	0.643	0.738	0.689
	CORSAIR	0.723	0.649	<u>0.744</u>	0.695
	GA	<u>0.727</u>	**0.653**	0.743	**0.710**
	EDGE	**0.731**	<u>0.650</u>	**0.750**	<u>0.706</u>

Table 2: *LoRA-Based Fine-Tuning* model Performance on Biased-SNLI. **OA** denotes Overall Accuracy, while **WGA** refers to Worst-Group Accuracy, with groups implicitly defined by the presence or absence of the injected spurious token and the class label. p_{prev} denotes bias prevalence, and p_{str} refers to bias strength, as described in Section 5.1.

5.1 Datasets Settings

We evaluate our method in two complementary settings to comprehensively assess its debiasing effectiveness and robustness:

Controlled Synthetic Bias Evaluation. To analyze model behavior under controlled spurious correlations, we adopt a synthetic bias-injection protocol inspired by prior work (Dranker et al., 2021; Joshi et al., 2022). We modify the SNLI dataset (Wang et al., 2018) by appending a spurious token t_s ("!!") to the hypothesis, inducing an artificial correlation with the target class c_k ("entailment"). The injection is governed by two parameters: **Bias prevalence** (p_{prev}): the probability that a sample is biased; and **Bias strength** (p_{str}): the conditional probability that the token appears in class c_k when a sample is biased. Mathematically, let $D = \{(x^i, y^i)\}_{i=1}^M$ denote the original dataset, where $x^i \in \mathcal{X}$ and $y^i \in \mathcal{Y} = \{c_1, \ldots, c_K\}$. The biased dataset $D_b = \{(\widetilde{x}^i, y^i)\}_{i=1}^M$ is generated as:

$$\widetilde{x}^i = \begin{cases} x^i, & \text{with probability } 1 - p_{\text{prev}}, \\ x^i \oplus t_s, & \text{with probability } p_{\text{prev}} \cdot q^i, \end{cases} \quad \text{where} \quad q^i = \begin{cases} p_{\text{str}}, & \text{if } y^i = c_k, \\ 1 - p_{\text{str}}, & \text{if } y^i \neq c_k, \end{cases} \tag{13}$$

Here, $\oplus$ denotes token concatenation. To induce spurious correlations, we configure the training and development sets with two bias settings: $(p_{\text{prev}}, p_{\text{str}}) \in \{(25\%, 90\%), (33\%, 50\%)\}$, as shown in Table 2. For the test set, we set $p_{\text{prev}} = 66\%$ and $p_{\text{str}} = 50\%$, creating a mismatch under distributional shift.

Real-World Robustness Evaluation. Beyond synthetic setups, we evaluate our method in real-world scenarios where spurious correlations naturally arise. In NLI tasks, for instance, linguistic artifacts such as negation frequently co-occur with specific labels (e.g., `Contradiction`) (Joshi et al., 2022; Zhang et al., 2024b), as illustrated in Figure 1. We conduct evaluations on two fronts. First, in the NLP domain, we fine-tune models on the original SNLI dataset (Wang et al., 2018) using LoRA adapters and assess robustness under both adversarial (Liu et al., 2020) and stress test conditions (Naik et al., 2018). To further examine cross-distribution generalization, we include MNLI-hard (matched and mismatched) (Mahabadi et al., 2019) and QNLI (Wang et al., 2018). A summary of the evaluation datasets is provided in Table 1. Second, to validate cross-modal generality, we evaluate in the vision domain using the Waterbirds dataset (Sagawa et al., 2019), a standard benchmark for studying spurious correlations in image classification.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 890-903, Jinan, China, August 11-14, 2025.

Backbone	Method	OOD Generalization			Robustness Evaluation		
		MNLI[1]	MNLI[2]	QNLI	Stress Test	HANS	IC-CS
RoBERTa-base	ERM	0.764	0.772	0.720	0.680	0.645	0.637
	REW	0.766	0.775	0.721	**0.688**	0.650	**0.641**
	LM	0.757	0.766	0.712	0.673	0.638	0.630
	CORSAIR	0.765	0.776	0.724	0.684	0.647	0.640
	GA	**0.770**	0.779	**0.725**	0.687	**0.653**	0.640
	EDGE	0.769	**0.780**	0.722	0.686	0.652	0.640
GPT2	ERM	0.706	0.713	0.659	0.610	0.585	0.578
	REW	**0.708**	0.717	**0.660**	**0.618**	0.590	0.581
	LM	0.699	0.708	0.652	0.604	0.580	0.572
	CORSAIR	0.705	0.716	0.658	0.615	0.595	0.579
	GA	0.705	0.710	0.656	0.617	**0.602**	**0.582**
	EDGE	**0.708**	**0.719**	0.658	0.616	0.600	0.579

Table 3: Generalization Performance Comparison on Real-World Generalization Benchmarks with *LoRA-Based Fine-Tuning* on SNLI. We evaluate both *out-of-distribution generalization* (left) and *robustness to adversarial/stress conditions* (right). MNLI[1] means refers to MNLI-hard-match, and MNLI[2] refers to MNLI-hard-mismatch.

Method	ERM	LfF	DFA	CNC	GA	EDGE
Overall Accuracy	0.8009	0.8149	0.8240	0.8710	**0.8814**	0.8812
Worst-group Accuracy	0.6223	0.6562	0.6910	0.8540	**0.8746**	0.8548

Table 4: Performance comparison on the Waterbirds dataset using ResNet-18 with *full fine-tuning*. Groups are defined by combinations of class labels (waterbird, landbird) and bias attributes (water background, land background).

5.2 Experiment Setup

Baselines. We compare EDGE with standard ERM and state-of-the-art debiasing methods. For NLP experiments, we include Reweighting (REW) (Clark et al., 2019), Learned-Mixin (LM) (Clark et al., 2019), CORSAIR (Qian et al., 2021), and Gradient Alignment (GA) (Zhao et al., 2024). For CV experiments, we additionally consider DFA (Lee et al., 2021) and CNC (Zhang et al., 2022).

Model Implementation. Hyperparameters are fine-tuned on the validation set, with early stopping applied for optimal selection. For NLP experiments, we adopt RoBERTa-base and GPT2 as backbone models, and apply *LoRA fine-tuning strategy*. The hyperparameters are set as follows: batch size of 32, learning rate of 3×10^{-5} with weight decay 0.05. Additionally, we apply a learning rate warm-up for 1000 steps. For CV experiments, we use ResNet-18 with *full fine-tuning strategy*. The hyperparameters are set as follows: batch size of 256, learning rate of 1×10^{-3} .

5.3 Main Results

Controlled Synthetic Bias Setting. Table 2 presents the debiasing performance of various methods across all backbone models. The results demonstrate that EDGE consistently achieves superior effectiveness and robustness, outperforming existing debiasing baselines under a range of bias configurations ($p_{\text{prev}}, p_{\text{str}}$). By identifying the common gradient direction shared by different training objectives and isolating the bias gradient component, EDGE enables explicit debiasing through gradient manipulation. This design allows it to generalize well even under distributional shifts with varying degrees of spurious correlations. Among the debiasing baselines, most methods (e.g., REW, LM, CORSAIR) adopt strategies

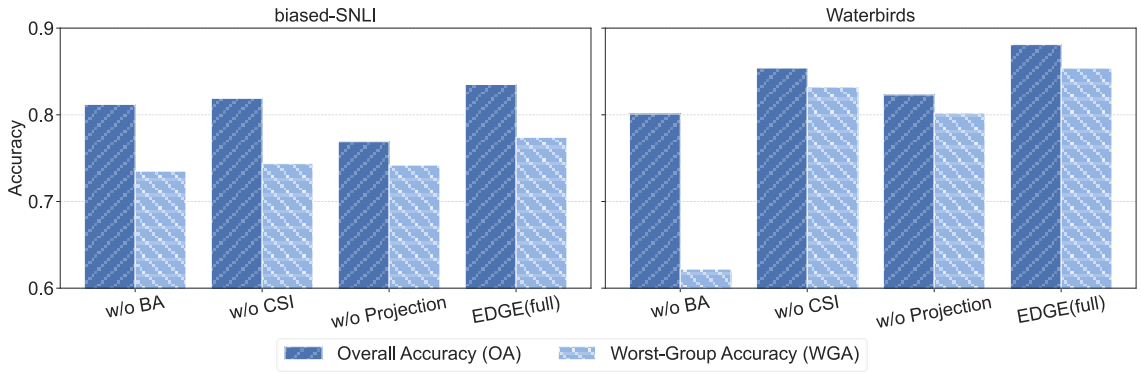

Figure 4: Ablation experiments conducted on Biased-SNLI (left) and Waterbirds (right). "BA" refers to Bias Amplification, and "CSI" refers to Common Subspace Identification.

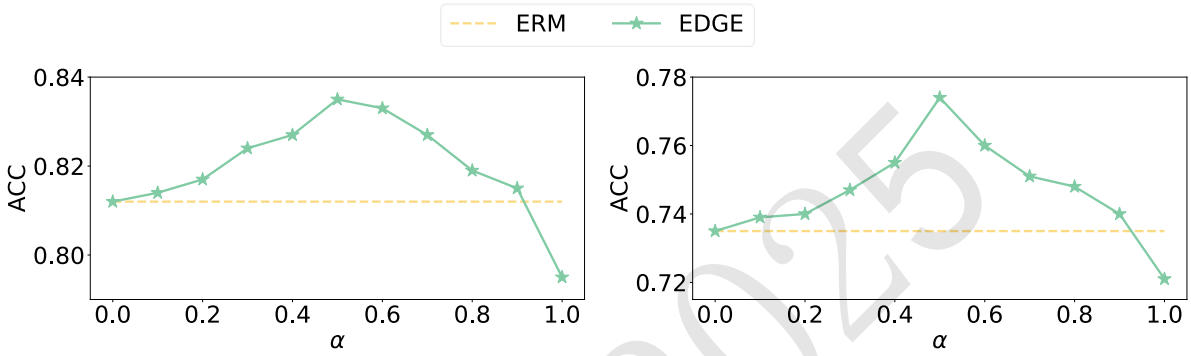

Figure 5: Parameter Sensitivity Test with Synthetically Biased SNLI Dataset on RoBERTa-base. (Left) Overall accuracy. (Right) Worst-Group Accuracy.

that emphasize atypical samples during training to mitigate bias. As a result, several of them exhibit competitive OOD performance, particularly GA, which leverages gradient alignment to enhance the influence of atypical examples. However, REW exhibits inconsistent performance and sometimes underperforms the backbone. We attribute this to its reliance on prediction confidence for weighting ($w_i = 1 - p_y$), which fails to account for the sparsity and distribution of atypical instances—particularly non-target samples ($y \neq c_k$) that contain spurious features (t_s). This can lead to suboptimal bias disentanglement and reduced robustness. In contrast, EDGE is both model-agnostic and fine-tuning-strategy-agnostic. Unlike LM or CORSAIR, it can be seamlessly applied under both full-parameter fine-tuning and PEFT regimes, making it highly versatile and scalable across different deployment scenarios.

Real-World Robustness Setting. To assess the generalization ability of EDGE under realistic distribution shifts, we evaluate its performance across diverse NLP and vision benchmarks, as summarized in Table 3 and Table 4. In the NLP domain, EDGE consistently improves robustness across multiple test conditions, demonstrating strong out-of-distribution generalization. In the vision domain, we evaluate on the Waterbirds dataset (Sagawa et al., 2019), where spurious correlations (e.g., background cues) are prevalent. EDGE consistently outperforms most existing debiasing baselines, demonstrating strong generalization without requiring explicit bias labels. Although both EDGE and GA can achieve comparable performance in some scenarios, GA relies on explicit bias labels and performs gradient alignment through group-specific reweighting during training, making it mainly suitable for cases with a limited number of predefined bias groups. In contrast, EDGE does not require access to bias labels and is inherently scalable to arbitrary group structures, thus offering a more principled and generalizable solution for balancing debiasing and generalization across diverse domains.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 890-903, Jinan, China, August 11-14, 2025.

Dataset	Sample	Label	ERM	**EDGE**
MNLI^2	P: It 's **not** a family story . H: It 's a story about drugs and murder .	N	C ✗	N ✓
	P: How old are you talking about ? H: Where are you talking about ?	C	E ✗	C ✓
HANS	P: The doctors were mentioned by the bankers. H: The bankers mentioned the doctors.	E	NE ✗	E ✓
	P: The scientist believed the artists ran . H: The scientist believed the artists .	NE	E ✗	NE ✓

Table 5: Case study on the MNLI^2 and HANS development sets using RoBERTa-base as the backbone. {'C', 'E', 'N'} denote Contradiction, Entailment, and Neutral; for HANS, 'NE' denotes Not Entailment.

5.4 Detailed Comparison and Analysis

Ablation Study. To better understand the contribution of each component in EDGE, we conduct a series of ablation experiments. Specifically, we evaluate the impact of (1) bias amplification, (2) common subspace identification, and (3) gradient projection, by progressively disabling or modifying each module. Under the synthetic biased SNLI setup using RoBERTa-base as the backbone. we report both Overall Accuracy (OA) and Worst-Group Accuracy (WGA) to assess generalization and debiasing effectiveness. As shown in Figure 4, removing any core component of EDGE causes significant performance degradation. Without bias amplification or subspace identification, the model fails to expose spurious features. Skipping orthogonal projection and naively subtracting the amplified gradient severely hampers task learning, resulting in the worst performance. These results highlight the critical role of each component in EDGE, showing that bias amplification, subspace identification, and careful projection are essential for effective debiasing while preserving task knowledge.

Parameter Sensitive Test. The hyper-parameter α in Eq. 12 controls the extent of debiasing. Thus, we conduct experiments to verify its effect and report results under controlled synthetic bias settings in Figure 5. As α increases, the debiasing strength becomes more aggressive. We observe that moderate values of α (e.g., around 0.5) strike the best balance—effectively suppressing bias-induced features while preserving essential task-relevant information. Overly large α can harm performance by excessively removing useful gradients, while small α fails to sufficiently mitigate bias. These findings validate the role of α as a tunable mechanism to navigate the trade-off between debiasing and model utility.

Case Study. We present a qualitative case study on MNLI^2 and HANS using RoBERTa-base to illustrate the behavioral differences between ERM and EDGE under distribution shifts (Table 5). For **MNLI^2**, we highlight two typical failure cases of ERM. In the first example, the presence of a negation token *"not"* misleads the ERM model into predicting `Contradiction`, revealing its reliance on lexical cues. In contrast, EDGE correctly identifies the relationship as `Neutral`, demonstrating robustness against spurious negation. In the second example, the high lexical overlap between premise and hypothesis causes ERM to predict `Entailment`, while EDGE accurately recognizes a subtle semantic inconsistency and predicts `Contradiction`. For **HANS**, we analyze two syntactic traps. In the first, a passive-to-active subject-object swap confuses the ERM model into predicting `Not Entailment`, whereas EDGE correctly preserves the entailment relation. The second case is more challenging: the hypothesis is a clause of the premise, leading ERM to over-rely on surface overlap and predict `Entailment`. However, EDGE successfully detects the lack of entailment due to ellipsis and structural mismatch, outputting `Not Entailment`. These examples highlight how ERM models tend to exploit easy-to-learn but non-robust features (e.g., negation tokens, word overlap), whereas EDGE mitigates such biases by refining the gradient signal, enabling more semantically grounded decisions under OOD settings.

6 Conclusion

In this work, we introduce EDGE, a simple yet effective debiasing framework that explicitly disentangles model gradients into bias-relevant and task-relevant components. By identifying a shared feature subspace across different training configurations, EDGE removes bias gradients in a principled and model-agnostic manner. Unlike prior methods that depend on confidence heuristics or specific fine-tuning schemes, EDGE is compatible with both full and PEFT, making it broadly applicable. Extensive experiments on synthetic and real-world benchmarks demonstrate that EDGE consistently enhances robustness and out-of-distribution generalization, establishing gradient-level disentanglement as a powerful paradigm for addressing dataset bias.

Acknowledgements

This research was partially supported by grants from National Science and Technology Major Project under Grant (No. 2023ZD0121103), the National Natural Science Foundation of China (No. 62376086), and the Fundamental Research Funds for the Central Universities of China (No. JZ2025HGTG0289, PA2025IISL0114).

References

Sumyeong Ahn, Seongyoon Kim, and Se-Young Yun. 2022. Mitigating dataset bias by using per-sample gradient. *arXiv preprint arXiv:2205.15704*.

AI@Meta. 2025. Llama 4 model card.

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

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

Saswat Das, Marco Romanelli, Cuong Tran, Zarreen Reza, Bhavya Kailkhura, and Ferdinando Fioretto. 2024. Low-rank finetuning for llms: A fairness perspective.

Yana Dranker, 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.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models. *arXiv preprint arXiv:2103.06922*.

Henry Gouk, Timothy M Hospedales, and Massimiliano Pontil. 2020. Distance-based regularisation of deep networks for fine-tuning. *arXiv preprint arXiv:2002.08253*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Taeuk Jang and Xiaoqian Wang. 2023. Difficulty-based sampling for debiased contrastive representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24039–24048.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. *arXiv preprint arXiv:2210.14011*.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. 2021. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. 2022. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yuxin Liao, Yonghui Yang, Min Hou, Le Wu, Hefei Xu, and Hao Liu. 2025. Mitigating distribution shifts in sequential recommendation: An invariance perspective.

Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. 2023. Biasadv: Bias-adversarial augmentation for model debiasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3832–3841.

Tianyu Liu, Xin Zheng, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020. An empirical study on model-agnostic debiasing strategies for robust natural language inference. *arXiv preprint arXiv:2010.03777*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Yougang Lyu, Piji Li, Yechang Yang, Maarten de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, and Zhaochun Ren. 2023. Feature-level debiased natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13353–13361.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2019. End-to-end bias mitigation by modelling biases in corpora. *arXiv preprint arXiv:1909.06321*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

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.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of nli models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6867–6874.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

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

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585.

Zhiqiang Shen, Zechun Liu, Jie Qin, Marios Savvides, and Kwang-Ting Cheng. 2021. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9594–9602.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwen3, April.

Junjiao Tian, Zecheng He, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, and Zsolt Kira. 2023. Trainable projected gradient method for robust fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7836–7845.

Junjiao Tian, Yen-Cheng Liu, James S Smith, and Zsolt Kira. 2024. Fast trainable projection for robust fine-tuning. *Advances in Neural Information Processing Systems*, 36.

Rishabh Tiwari and Pradeep Shenoy. 2023. Overcoming simplicity bias in deep networks using a feature sieve. In *International Conference on Machine Learning*, pages 34330–34343. PMLR.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. *arXiv preprint arXiv:2005.00315*.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2022. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu, et al. 2023. Aprompt: Attention prompt tuning for efficient adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9147–9160.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. 2024. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24502–24511.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022a. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022b. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. *arXiv preprint arXiv:2203.12942*.

Sishi Xiong, Yu Zhao, Jie Zhang, Li Mengxiang, Zhongjiang He, Xuelong Li, and Shuangyong Song. 2024. Dual prompt tuning based contrastive learning for hierarchical text classification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12146–12158.

LI Xuhong, Yves Grandvalet, and Franck Davoine. 2018. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pages 2825–2834. PMLR.

Yonghui Yang, Le Wu, Yuxin Liao, Zhuangzhuang He, Pengyang Shao, Richang Hong, and Meng Wang. 2025. Invariance matters: Empowering social recommendation via graph invariant learning. *arXiv preprint arXiv:2504.10432*.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. 2022. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Dacao Zhang, Kun Zhang, Le Wu, Mi Tian, Richang Hong, and Meng Wang. 2024a. Path-specific causal reasoning for fairness-aware cognitive diagnosis. *arXiv preprint arXiv:2406.03064*.

Kun Zhang, Dacao Zhang, Le Wu, Richang Hong, Ye Zhao, and Meng Wang. 2024b. Label-aware debiased causal reasoning for natural language inference. *AI Open*, 5:70–78.

Dacao Zhang, Fan Yang, Kun Zhang, Xin Li, Si Wei, Richang Hong, and Meng Wang. 2025a. Optimizing low-rank adaptation with decomposed matrices and adaptive rank allocation. *Frontiers of Computer Science*, 19(5):195337.

Dacao Zhang, Kun Zhang, Shimao Chu, Le Wu, Xin Li, and Si Wei. 2025b. More: A mixture of low-rank experts for adaptive multi-task learning. *arXiv preprint arXiv:2505.22694*.

Kun Zhang, Le Wu, Kui Yu, Guangyi Lv, and Dacao Zhang. 2025c. Evaluating and improving robustness in large language models: A survey and future directions. *arXiv preprint arXiv:2506.11111*.

Bowen Zhao, Chen Chen, Qian-Wei Wang, Anfeng He, and Shu-Tao Xia. 2024. Delving into identify-emphasize paradigm for combating unknown bias. *International Journal of Computer Vision*, 132(6):2310–2330.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. 2023a. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. 2023b. Debiased fine-tuning for vision-language models by prompt regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3834–3842.