CONDITIONALLY MINIMAL SUFFICIENT REPRESENTATIONS FOR DEBIASED VISUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep visual models often exploit spurious correlations in training data, degrading performance on bias-conflicting groups. Many representation-level debiasing methods reduce global dependence between representations and bias variables, which can be overly restrictive when the target and bias variables are statistically dependent. To address this limitation, we propose **Conditionally Minimal Sufficient Debiasing Representation (CMDR)**, a conditional minimal-sufficiency framework that replaces global bias invariance with target-conditioned debiasing while limiting target-conditioned input redundancy. We theoretically characterize this formulation by establishing a feasibility separation between global invariance and conditional minimal sufficiency under a realizable setting, showing that $I(H; Z_b \mid Y)$ bounds target-conditioned group-risk gaps, and that $I(H; X \mid Y)$ bounds dependence on input-mediated nuisances under the corresponding Markov condition. CMDR instantiates these principles with a target prediction objective, a target-conditioned adversarial discriminator, and a lightweight variational regularizer, while supporting both explicit bias supervision and confidence-weighted pseudo-bias signals when training bias annotations are unavailable. Experiments on Waterbirds, CelebA, BAR, MetaShift, and BFFHQ demonstrate improved worst-group robustness across diverse visual biases and bias-conflicting ratios. Ablations show that target-conditioned debiasing outperforms global adversarial debiasing, while conditional minimality provides additional worst-group gains and confidence weighting improves robustness under weak supervision.

1 INTRODUCTION

Deep visual models often exploit spurious cues that correlate with target labels in the training data (Chen et al., 2023; Dehdashtian et al., 2024). Although such shortcuts can yield high in-distribution accuracy, they may fail under distribution shifts and induce substantial performance disparities across subgroups (Riaz & Smeaton, 2025; Aylapuram et al., 2025). For example, perception systems can become sensitive to changes in illumination or weather (Riaz & Smeaton, 2025), while face recognition models may exhibit accuracy gaps across demographic groups (Aylapuram et al., 2025). Such failures suggest that spurious correlations harm both out-of-distribution generalization and subgroup equity, raising concerns about robustness and fairness (Dehdashtian et al., 2024). As visual models (Tang et al., 2022; Zhang et al., 2022) are increasingly deployed in high-risk domains, learning representations that preserve target semantics while suppressing biased cues has become a critical challenge.

Existing approaches mitigate spurious correlations through data intervention, training-time regularization, or post-hoc correction (Aylapuram et al., 2025; Lin et al., 2024). Data-level methods adjust the training distribution through resampling, augmentation, or fairness-aware generation (Lin et al., 2024); training-time methods use reweighting, fairness constraints, or adversarial objectives to suppress shortcut learning (Aylapuram et al., 2025); and post-hoc methods remove learned bias signals through fine-tuning, representation calibration, or machine unlearning (Aylapuram et al., 2025). Despite their progress, these approaches often depend on prior knowledge of bias types, group annotations, or extra computation (Aylapuram et al., 2025; Lin et al., 2024). More fundamentally, representation-level debiasing often aims to suppress bias-related information globally, without distinguishing bias dependence explained by the target from residual shortcut information that remains within each target class.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A deeper issue arises from the objective itself. Many representation-level methods implicitly pursue global bias invariance by reducing the dependence between the representation H and the bias variable Z_b (Oh et al., 2022; Tian et al., 2024; Lin et al., 2023; Aylapuram et al., 2025). However, when Y and Z_b are statistically dependent, global invariance can conflict with retaining sufficient target information. We use $\mathcal{H}(\cdot)$ for entropy and $I(\cdot;\cdot)$ for mutual information. We show that

$$I(H; Z_b) \geq I(Y; Z_b) - \mathcal{H}(Y \mid H). \tag{1}$$

In particular, when $\mathcal{H}(Y \mid H) < I(Y; Z_b)$, global bias independence is impossible. Under $\mathcal{H}(Y \mid X) = 0$ and $I(Y; Z_b) > 0$, this further yields a feasibility separation: target sufficiency and exact global invariance are incompatible, while the three CMDR criteria remain jointly feasible. This motivates suppressing only the residual bias dependence conditioned on the target, $I(H; Z_b \mid Y)$, rather than $I(H; Z_b)$. Conditional debiasing alone, however, addresses only the specified bias variable and may leave other nuisance information in the representation. We therefore additionally limit $I(H; X \mid Y)$ to control input-mediated nuisance information beyond the modeled bias.

Motivated by this feasibility separation, we propose **Conditionally Minimal Sufficient Debiasing Representation (CMDR)**, a unified framework that jointly promotes target sufficiency, conditional debiasing, and conditional minimality. CMDR targets residual bias dependence $I(H; Z_b \mid Y)$ rather than global dependence, while regularizing $I(H; X \mid Y)$ to limit residual input information unexplained by the target. We further establish that conditional bias dependence bounds target-conditioned group-risk gaps, whereas $I(H; X \mid Y)$ upper-bounds dependence on any input-mediated nuisance satisfying the corresponding Markov condition. These principles are instantiated through a target prediction objective, a target-conditioned adversarial discriminator, and a lightweight variational conditional minimality regularizer. CMDR supports both explicit bias supervision and confidence-weighted pseudo-bias signals when training bias annotations are unavailable. Experiments across multiple visual bias benchmarks demonstrate consistent improvements in worst-group robustness, with ablations validating the complementary roles of the proposed components.

Our main contributions are as follows:

- We establish an information-theoretic feasibility separation between global invariance and conditional minimal sufficiency: under a realizable setting with statistically dependent target and bias variables, exact global invariance is incompatible with target sufficiency, while the CMDR criteria can remain jointly feasible. This motivates replacing $I(H; Z_b)$ with residual conditional dependence $I(H; Z_b \mid Y)$.
- We develop CMDR, a conditional minimal-sufficiency framework that jointly preserves target information, suppresses specified residual bias, and limits unspecified input-mediated nuisances. We further connect $I(H; Z_b \mid Y)$ to target-conditioned group-risk gaps and show that $I(H; X \mid Y)$ controls input-mediated nuisance dependence under the corresponding Markov condition. The framework supports both explicit bias labels and confidence-weighted pseudo-bias signals.
- We evaluate CMDR on five visual bias benchmarks across multiple bias types and bias-conflicting sample ratios. CMDR consistently improves worst-group robustness, while extensive ablations validate conditional debiasing, conditional minimality, and weakly supervised bias discovery.

2 Related Work

2.1 Debiasing under Spurious Correlations

Visual models often rely on spurious correlations between target and bias attributes, leading to poor worst-group performance under distribution shifts. Existing debiasing methods typically intervene at the data, training, or post-hoc stage. Data-level methods rebalance, select, or augment samples through resampling (Wang et al., 2020; Buda et al., 2018), data selection (Jain et al., 2024), data or feature augmentation (Lim et al., 2023; Jung et al., 2025), or fairness-aware generation (Lin et al., 2024). Training-time methods use group-robust objectives (Sagawa et al., 2020; Słowik & Bottou, 2022), failure-based reweighting or resampling (Liu et al., 2021; Nam et al., 2020; Han et al., 2024), bias-aware sampling (Qraitem et al., 2023), or representation regularization (Park et al., 2024). Post-hoc methods adjust pretrained models or representations through calibration (You et al.,

2024), fine-tuning (Kirichenko et al., 2022; Zhu et al., 2025), unlearning (Aylapuram et al., 2025), or feature-space correction (Li et al., 2025). CMDR formulates debiasing at the representation level by distinguishing bias dependence explained by the target from residual bias dependence that remains within each target class.

2.2 Adversarial and Conditional Representation Debiasing

Fair and adversarial representation learning aims to learn task-relevant features that are insensitive to protected, domain, or bias variables (Zemel et al., 2013; Ganin et al., 2016; Zhang et al., 2018; Kim et al., 2019). Recent visual fairness methods further mitigate bias through disentanglement (Oh et al., 2022; Lin et al., 2023), separability constraints (Jang & Wang, 2024; Jang et al., 2024), attention debiasing (Lu et al., 2024), or architecture-specific regularization (Tian et al., 2024). Many adversarial or invariance-based methods can be viewed as reducing global dependence between the representation H and the bias variable Z_b, i.e., $I(H; Z_b)$.

Global invariance can be overly restrictive when the target and bias variables are statistically dependent. Conditional invariance has been studied in fairness, domain adaptation, and domain generalization (Hardt et al., 2016; Long et al., 2018; Pham et al., 2023). Minimal-representation and information-bottleneck formulations seek compact representations that preserve task-relevant information while reducing input redundancy, with conditional formulations using $I(H; X \mid Y)$ to quantify target-conditioned residual information (Fischer, 2020; Tezuka & Namekawa, 2021). CMDR differs in its objective formulation: it treats $I(H; Z_b \mid Y)$ as the residual bias dependence to be suppressed and couples it with conditional minimality through $I(H; X \mid Y)$. This yields a conditional minimal-sufficiency objective that promotes target sufficiency, targets specified residual bias, and limits unspecified input-mediated nuisances within a unified representation-learning framework.

2.3 Debiasing with Weak or Missing Bias Supervision

Many group-robust and debiasing methods rely on explicit group or bias annotations, which may be expensive or unavailable. Recent work therefore infers pseudo-groups (Han et al., 2024), pseudo-attributes (Seo et al., 2022), or calibration signals (You et al., 2024; Zhu et al., 2025; Li et al., 2025) from learned representations. Other methods adapt pretrained vision-language models through test-time or parameter-efficient correction without full retraining (Lu et al., 2025; Gerych et al., 2024). These approaches reduce annotation requirements, but often use inferred bias information for reweighting, resampling, calibration, or post-hoc adjustment.

CMDR also supports missing bias supervision, but uses pseudo-bias information as soft nuisance supervision inside the same conditional representation objective. When bias labels are available, Z_b is observed; otherwise, CMDR estimates confidence-weighted pseudo-bias directions from frozen visual features. This allows the same conditional minimal-sufficiency framework to operate under both explicit bias supervision and training-time group-agnostic settings.

3 Method

We propose CMDR, a conditional minimal-sufficiency framework for visual debiasing. Motivated by a feasibility separation between global invariance and conditional minimal sufficiency, CMDR combines target-conditioned debiasing with conditional minimality and supports both explicit bias labels and confidence-weighted pseudo-bias signals.

3.1 Problem Setup

Let the training set be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathcal{X}$ is an input image and $y_i \in \{1, \dots, C\}$ is its target. Each sample may additionally be associated with a bias variable z_i^b. It can be an observed bias attribute, or a multidimensional latent bias signal discovered from pretrained features when training bias labels are unavailable. In the weakly supervised setting, each pseudo-bias estimate is assigned a confidence weight $w_i \in [0, 1]$; with explicit bias labels, we set $w_i = 1$.

We learn a deterministic feature $h = \phi_\theta(X)$ and classifier f_ψ that preserve target information while suppressing within-class bias and nuisance information. For the information-theoretic formula-

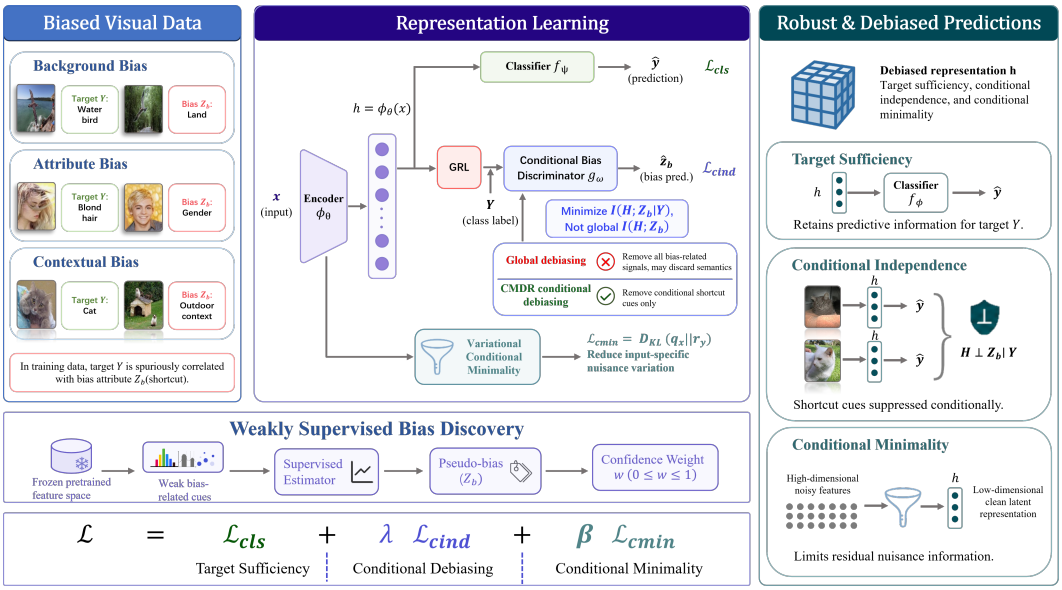

Figure 1: Overview of CMDR. CMDR jointly optimizes target sufficiency $\mathcal{L}_{\text{cls}}$, target-conditioned debiasing $\mathcal{L}_{\text{cind}}$, and variational conditional minimality $\mathcal{L}_{\text{cmin}}$ based on $D_{\text{KL}}(q_x \| r_y)$, where $q_x = q_\theta(H \mid x)$ and $r_y = r_\eta(H \mid y)$. Pseudo-bias signals and confidence weights are used when training bias labels are unavailable.

tion, H denotes the representation random variable induced by X. For the tractable conditional-minimality regularizer introduced below, we instantiate $q_\theta(H \mid x)$ as a fixed-variance Gaussian centered at h. Prediction and adversarial debiasing operate on the mean feature h.

We first characterize a fundamental limitation of global debiasing. For any representation H,

$$I(H; Z_b) \geq I(Y; Z_b) - \mathcal{H}(Y \mid H). \tag{2}$$

Thus, if $\mathcal{H}(Y \mid H) < I(Y; Z_b)$, exact global bias independence is impossible. Under $\mathcal{H}(Y \mid X) = 0$ and $I(Y; Z_b) > 0$, this yields a feasibility separation: target sufficiency and global invariance are incompatible, while the three CMDR criteria remain jointly feasible.

We next characterize the role of conditional minimality beyond a specified bias variable. For any input-mediated nuisance U satisfying $U \perp H \mid X, Y$,

$$I(H; U \mid Y) \leq I(H; X \mid Y). \tag{3}$$

Hence, minimizing $I(H; X \mid Y)$ simultaneously limits dependence on unspecified input-mediated nuisances. Together, conditional debiasing controls the modeled bias Z_b, while conditional minimality controls residual nuisances beyond Z_b. For discrete groups, Appendix further derives a bound connecting $I(H; Z_b \mid Y)$ to target-conditioned group-risk gaps.

These characterizations lead to the following conditional minimal-sufficiency objective:

$$\min_\theta I(Y; X \mid H) + \lambda I(H; Z_b \mid Y) + \beta I(H; X \mid Y). \tag{4}$$

The three terms measure lack of target sufficiency, residual conditional bias, and target-conditioned input redundancy, respectively. Unlike global invariance, this objective explicitly separates target information, specified residual bias, and unspecified nuisance variation. Detailed derivations and theoretical guarantees are provided in Appendix.

3.2 CMDR Method

Overall framework. CMDR instantiates the three information-theoretic terms in Eq. 4 with tractable classification, adversarial, and variational objectives. It uses an encoder ϕ_θ, target predictor

f_ψ, training-time target-conditioned bias discriminator g_ω, and lightweight target-conditioned variational prior r_η (Figure 1). Given x_i, the encoder produces $h_i = \phi_\theta(x_i)$ and the predictor estimates $\hat{y}_i = f_\psi(h_i)$. During training, $g_\omega(h_i, y_i)$ guides the encoder to suppress recoverable conditional bias, while r_η regularizes the representation toward target-conditioned minimality. The discriminator is discarded at inference.

Sufficiency via classification. To preserve target-relevant information, we use the standard cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N} \text{CE}\big(f_\psi(\phi_\theta(x_i)),\, y_i\big). \tag{5}$$

This loss provides a tractable surrogate for target insufficiency, encouraging the learned representation to retain information required for predicting Y.

Target-conditioned debiasing. Motivated by the incompatibility in Eq. 2, CMDR replaces global bias invariance with target-conditioned dependence minimization:

$$I(H; Z_b \mid Y), \tag{6}$$

which targets residual within-class bias without suppressing dependence explained by Y.

We suppress this dependence through a target-conditioned bias discriminator g_ω applied to the mean feature h, from which the stochastic representation H is constructed. Given h_i and target label y_i, the discriminator predicts

$$\hat{z}_i^b = g_\omega(h_i, y_i). \tag{7}$$

The conditional debiasing loss is

$$\mathcal{L}_{\text{cind}} = \frac{1}{N}\sum_{i=1}^{N} w_i\, \ell_b\big(g_\omega(\text{GRL}(h_i), y_i),\, z_i^b\big), \tag{8}$$

where ℓ_b is cross-entropy for discrete z_i^b and mean squared error for continuous z_i^b. We set $w_i = 1$ with explicit bias labels; otherwise, it down-weights unreliable pseudo-bias estimates.

For discrete Z_b, under an optimal conditional discriminator, the cross-entropy loss estimates $\mathcal{H}(Z_b \mid h, Y)$. Through the GRL, the encoder adversarially increases this loss, thereby reducing $I(h; Z_b \mid Y)$ since $\mathcal{H}(Z_b \mid Y)$ is fixed. For continuous pseudo-biases, we use MSE as a practical target-conditioned predictability surrogate, without claiming exact conditional mutual-information minimization.

Variational conditional minimality regularization. Conditional debiasing targets the specified Z_b but does not constrain unmodeled nuisances. By Eq. 3, controlling $I(H; X \mid Y)$ bounds the conditional dependence on any input-mediated nuisance U. We therefore variationally regularize $I(H; X \mid Y)$.

Specifically, we define the stochastic representation and target-conditioned variational prior as

$$q_\theta(H \mid x_i) = \mathcal{N}(H; h_i, \tau^2 I), \qquad r_\eta(H \mid y_i) = \mathcal{N}\left(H; m_{y_i}, \text{diag}(s_{y_i}^2)\right), \qquad s_{y_i}^2 > 0. \tag{9}$$

The conditional minimality loss is

$$\mathcal{L}_{\text{cmin}} = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}\left(q_\theta(H \mid x_i) \,\|\, r_\eta(H \mid y_i)\right). \tag{10}$$

Here $r_\eta(H \mid y_i)$ is jointly learned with class-specific mean and diagonal variance, while τ^2 is fixed. The expected KL objective is a variational upper bound on $I(H; X \mid Y)$, thereby limiting input information unexplained by Y.

Final objective. Combining all components, CMDR optimizes

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{cind}} + \beta \mathcal{L}_{\text{cmin}}, \tag{11}$$

where λ and β control conditional debiasing and minimality, respectively. This objective provides a tractable realization of Eq. 4: the three terms respectively promote target sufficiency, residual-bias suppression, and input-mediated nuisance control.

Table 1: **Group robustness of different methods on Waterbirds, CelebA, and BAR.** We report worst-group accuracy (WGA, primary) and Avg for CLIP ResNet-50 and ViT-L/14. Bold and underlined entries denote the best and second-best WGA within each supervision setting, respectively.

Method	CLIP ResNet-50						CLIP ViT-L/14					
	Waterbirds		CelebA		BAR		Waterbirds		CelebA		BAR	
	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg
Methods without training-time subgroup supervision												
ERM	43.769	96.120	47.778	97.954	45.070	75.027	57.009	97.345	28.889	97.822	77.465	96.146
JTT	61.994	94.164	56.667	97.468	<u>56.338</u>	78.648	57.788	97.316	42.222	97.662	74.648	95.294
LfF	63.459	64.712	17.478	38.747	55.704	79.057	63.240	97.323	**87.253**	89.092	78.873	96.376
BE	<u>67.134</u>	92.039	63.387	75.763	52.887	72.785	71.495	97.427	41.667	97.683	<u>80.282</u>	96.211
CFR	61.994	93.780	<u>63.889</u>	97.087	54.798	75.677	69.159	97.127	70.636	96.075	77.465	96.146
NSF	55.432	93.820	57.778	97.627	54.547	75.327	67.757	97.272	<u>83.333</u>	93.391	77.465	95.699
TAB	62.773	95.548	62.778	94.678	55.648	73.294	60.592	96.756	61.165	87.915	74.648	95.478
Zero-Shot	54.200	92.400	55.000	88.000	53.100	70.500	26.500	88.200	27.000	85.900	73.800	93.700
TTR	-	-	-	-	-	-	<u>74.140</u>	87.210	83.330	86.000	77.560	95.440
AGS	63.282	93.156	53.889	97.918	55.556	71.298	70.156	97.122	77.222	97.750	75.211	94.030
CMDR-Pseudo	**70.405**	93.483	**77.778**	96.035	**57.746**	73.081	**83.178**	97.777	78.667	93.878	**83.099**	96.695
Methods with subgroup supervision												
BM	<u>73.836</u>	85.191	85.103	92.340	<u>58.739</u>	72.454	63.551	84.566	80.000	93.260	63.380	93.853
GroupDRO	67.601	91.436	86.667	93.237	58.113	79.910	88.474	96.274	88.333	92.819	80.282	96.395
OS	54.191	87.733	82.222	93.146	51.481	64.644	87.850	96.295	87.560	91.682	77.465	95.962
UW	68.536	83.828	<u>87.778</u>	89.172	52.778	72.902	89.091	95.417	<u>88.676</u>	91.796	78.873	95.982
DFR	73.200	83.800	80.000	92.800	57.300	83.900	89.700	97.800	85.600	90.800	79.400	95.800
Hierarchical DRO	69.047	89.047	<u>87.778</u>	91.120	56.338	70.730	<u>90.745</u>	92.196	88.444	91.434	**87.324**	94.818
CMDR-Bias	**74.455**	94.134	**88.045**	93.152	**59.155**	72.687	**90.810**	95.560	**88.816**	91.070	<u>84.507</u>	97.293

Weakly supervised bias discovery. When bias labels are unavailable, we estimate z_i^b and confidence w_i in a frozen pretrained feature space. The pseudo-bias need not recover the ground-truth bias attribute; it provides soft nuisance supervision for conditional debiasing. We construct weak cues from image statistics, style responses, and augmentation-induced feature differences, and use partial least squares (PLS) regression to obtain low-dimensional continuous pseudo-biases. Confidence combines reconstruction consistency, signal strength, and estimation stability. No training bias or group annotations are used for pseudo-bias discovery. With explicit bias labels, we use the true labels and set $w_i = 1$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CMDR on five visual bias benchmarks covering background, attribute, and contextual biases. **Waterbirds** (Sagawa et al., 2020) is used to study background shortcuts, where waterbird/landbird labels are strongly correlated with water/land backgrounds. **CelebA** (Liu et al., 2015) evaluates attribute co-occurrence bias, using *blond hair* as the target attribute and *gender* as the bias attribute. **BAR** (Nam et al., 2020) focuses on contextual bias in action recognition, where action categories often co-occur with specific scene attributes. **MetaShift (Cat vs. Dog)** (Liang et al., 2022) further tests contextual distribution shifts under different bias-conflict ratios, with indoor/outdoor contexts correlated with class labels. **BFFHQ** (Kim et al., 2021) is used for facial bias evaluation, where *age* is the target attribute and *gender* is the bias attribute; we report results under 0.5%, 1%, 2%, and 5% bias-conflicting sample ratios. For all datasets, we follow the standard train/validation/test splits. For methods without training-time subgroup supervision, group annotations are not used for training or pseudo-bias discovery. Following common practice, validation group annotations are used only for model selection, and test group annotations only for computing WGA. Bias-supervised methods use explicit bias or group labels during training and are evaluated under the same splits and training budget.

Evaluation Metrics. We use **worst-group accuracy (WGA)** as the primary metric, i.e., the minimum accuracy across groups defined by the target label and bias attribute. Average accuracy (Avg) is also reported to reflect overall performance and its trade-off with group robustness. Unless otherwise specified, WGA is the main metric and Avg is secondary.

Table 2: **Group robustness under different bias-conflicting sample ratios on MetaShift and BFFHQ.** We report worst-group accuracy (WGA, primary) and Avg on CLIP ResNet-50 under different proportions of bias-conflicting training samples. Bold and underlined entries denote the best and second-best WGA within each supervision setting, respectively.

Method	MetaShift				BFFHQ							
	1%		5%		0.5%		1%		2%		5%	
	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg
Methods without training-time subgroup supervision												
ERM	77.263	99.200	90.157	98.918	56.8	99.2	63.2	84.7	72.8	89.0	86.0	94.2
BE	76.896	98.169	88.427	97.111	**64.0**	96.8	69.6	84.7	77.6	88.4	84.0	90.2
CFR	79.365	99.256	<u>91.946</u>	99.378	60.4	83.7	70.8	85.2	75.6	88.5	87.2	94.7
JTT	79.912	98.436	91.139	98.358	<u>63.6</u>	99.4	68.4	84.0	76.4	88.4	88.0	93.7
LfF	<u>81.015</u>	93.177	89.150	91.580	62.8	98.4	63.6	77.2	63.6	65.9	80.0	89.8
NSF	51.675	98.375	91.887	99.378	54.0	82.4	70.4	85.5	70.8	88.6	82.8	94.2
TAB	74.614	98.279	90.157	97.807	55.6	83.4	65.6	84.6	71.6	85.1	85.6	92.4
AGS	80.132	99.022	91.320	98.936	58.8	81.0	<u>71.2</u>	85.5	<u>78.4</u>	89.2	<u>89.2</u>	94.5
CMDR-Pseudo	**82.540**	99.394	**93.490**	98.993	**64.0**	84.4	**76.0**	88.0	**82.4**	91.4	**90.0**	95.1
Methods with subgroup supervision												
GroupDRO	<u>82.716</u>	96.392	<u>93.736</u>	95.801	<u>68.4</u>	98.8	75.6	87.0	83.6	91.5	90.8	94.7
OS	77.778	96.485	90.416	94.946	57.6	95.4	65.2	83.0	75.6	88.3	85.6	91.7
UW	79.470	96.298	92.405	97.336	67.2	91.8	**78.0**	88.8	**84.8**	91.0	**93.2**	94.7
DFR	81.789	96.452	92.465	96.254	66.8	98.4	74.9	88.6	82.4	90.6	89.4	93.5
Hierarchical DRO	82.187	87.072	92.767	95.249	67.6	84.7	74.8	88.4	83.6	90.9	92.0	95.4
CMDR-Bias	**84.106**	98.905	**94.407**	98.932	**72.0**	99.4	<u>76.8</u>	88.6	<u>84.4</u>	92.1	<u>92.8</u>	96.2

Baselines. We compare CMDR with representative visual debiasing and group-robust learning methods under two settings. First, we report ERM as the standard baseline. In the setting without training-time subgroup supervision, we compare with JTT 2021, LfF 2020, BE 2023, CFR 2024, NSF 2025, TAB 2024, Zero-Shot CLIP 2021, TTR 2025, and AGS 2026. We compare these methods with CMDR-Pseudo, which uses weakly discovered pseudo-bias signals without explicit bias labels. In the setting with subgroup supervision, we compare with Over-Sampling (OS) 2020, Upweighting (UW) 2019, BM 2023, GroupDRO 2020, DFR 2022, and Hierarchical DRO 2026, all of which use subgroup or bias annotations during training. We compare this group with CMDR-Bias, which uses the true bias labels in the target-conditioned adversarial objective. Following most prior work, we use test-set group annotations only to compute WGA for evaluation. Whenever applicable, methods are evaluated on the same train/validation/test splits with CLIP ResNet-50 and CLIP ViT-L/14 backbones. Further implementation details are provided in the Supplementary Material.

4.2 Main Results

Tables 1 and 2 report the main results on five visual bias benchmarks. We compare **CMDR-Pseudo** with methods without training-time subgroup supervision and **CMDR-Bias** with methods using subgroup supervision. Overall, CMDR achieves strong worst-group accuracy (WGA) across different bias types, model architectures, and bias-conflicting ratios.

Without training-time subgroup supervision. Without explicit bias labels, CMDR-Pseudo consistently improves over ERM across all five benchmarks and remains among the strongest methods without training-time subgroup supervision. Importantly, the improvement persists across background, attribute, and contextual biases, as well as across different model architectures and bias-conflicting ratios. This suggests that the discovered pseudo-bias signals provide useful nuisance supervision across different shortcut types, allowing the same conditional debiasing formulation to operate without manually specified bias annotations.

With subgroup supervision. When explicit bias or group annotations are available, CMDR-Bias remains among the leading methods across all five benchmarks. The strong performance of CMDR under both pseudo and explicit bias supervision indicates that its effectiveness is not tied to the particular weak bias-discovery procedure. Instead, the same conditional minimal-sufficiency formulation can incorporate more reliable bias information when available, while remaining applicable across different bias types, model architectures, and bias-conflicting ratios.

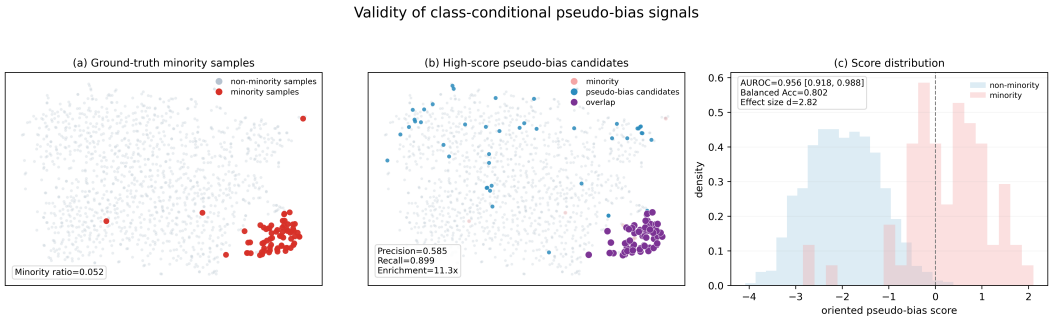

Figure 2: Validity of pseudo-bias signals. We visualize an example target class from the CelebA training split using t-SNE embeddings of model features. Precision, recall, and enrichment measure how well high-score pseudo-bias candidates recover minority samples, while AUROC, balanced accuracy, and effect size measure the separability of the pseudo-bias score. Ground-truth minority annotations are used only for analysis.

Trade-off between worst-group and average performance. We further examine whether improved worst-group robustness comes at the expense of overall accuracy. Across most settings, CMDR provides substantial WGA gains while retaining competitive Avg, suggesting that the improvement in worst-group robustness is generally achieved without severely degrading overall performance. A more visible trade-off appears in some highly biased regimes with extremely scarce bias-conflicting samples, particularly under pseudo-bias supervision. Overall, these results indicate that CMDR achieves a favorable balance between worst-group robustness and average predictive performance across diverse settings.

Robustness under scarce bias-conflicting samples. Table 2 further examines whether CMDR depends on having sufficient bias-conflicting examples during training. Across MetaShift and BFFHQ, the advantage of CMDR persists as the proportion of bias-conflicting samples decreases, including highly biased regimes where such examples are extremely scarce. Notably, CMDR-Pseudo retains this behavior without training-time subgroup annotations, suggesting that its weak bias supervision does not require abundant bias-conflicting examples to remain useful.

4.3 Ablation Study

Analysis of Pseudo-bias Discovery. Figure 2 evaluates the quality of pseudo-bias discovery without bias supervision. For visualization, we randomly sample 5,000 images from the CelebA training split, extract model features, and project them with t-SNE. In this example class, minority samples form a localized region in the target-conditioned feature space. Although they account for only 5.2% of samples, high-score pseudo-bias candidates achieve 58.5% precision, 89.9% recall, and 11.3$\times$ enrichment over random selection. The oriented pseudo-bias score also separates minority and non-minority samples with an AUROC of 0.956, balanced accuracy of 0.802, and effect size $d = 2.82$. These results suggest that the discovered pseudo-bias provides an informative soft bias direction for conditional debiasing, rather than a hard reconstruction of bias labels.

Effect of each component. Table 3 reports the contribution of each CMDR component. Minimality regularization alone brings limited gains, showing that penalizing input-specific redundancy is insufficient for debiasing. In contrast, the *target-conditioned* objective consistently outperforms global adversarial debiasing on Waterbirds, CelebA, and BAR, indicating that target-conditioned bias suppression is more effective than global adversarial bias suppression. Combining it with minimality regularization achieves the best WGA, as the former suppresses target-conditioned bias cues and the latter limits target-conditioned input redundancy and residual nuisance information.

Weakly supervised design choices. In weak supervision, noisy pseudo-bias variables make training stability crucial. Table 4 shows that global adversarial training with weak bias signals improves

Table 3: **Component ablation of CMDR-Bias on standard group-robustness benchmarks.** Results are reported on CLIP ViT-L/14. Global Adv. replaces the target-conditioned adversary with a global adversary without conditioning on Y. Cond. Adv. denotes our target-conditioned adversarial objective. Bold indicates the best WGA, and underlined denotes the second best WGA in each column.

Method	Waterbirds		CelebA		BAR	
	WGA↑	Avg	WGA↑	Avg	WGA↑	Avg
ERM	$57.009_{\pm 3.3}$	$97.345_{\pm 0.2}$	$28.889_{\pm 0.1}$	$97.822_{\pm 0.1}$	$77.465_{\pm 2.8}$	$96.146_{\pm 0.6}$
+ $\mathcal{L}_{\text{cmin}}$	$58.214_{\pm 2.4}$	$97.281_{\pm 0.3}$	$30.476_{\pm 0.1}$	$97.705_{\pm 0.1}$	$77.982_{\pm 2.3}$	$96.031_{\pm 0.6}$
+ Global Adv.	$87.214_{\pm 0.5}$	$95.031_{\pm 0.4}$	$86.941_{\pm 0.1}$	$91.204_{\pm 1.6}$	$78.864_{\pm 1.3}$	$96.021_{\pm 0.5}$
+ Cond. Adv.	$\underline{88.864_{\pm 0.6}}$	$94.991_{\pm 0.8}$	$\underline{87.311_{\pm 0.3}}$	$90.812_{\pm 1.1}$	$\underline{80.607_{\pm 1.6}}$	$96.508_{\pm 0.5}$
CMDR-Bias (full)	$\mathbf{90.810}_{\pm 0.2}$	$95.560_{\pm 1.3}$	$\mathbf{88.816}_{\pm 0.5}$	$91.070_{\pm 1.3}$	$\mathbf{84.507}_{\pm 1.4}$	$97.293_{\pm 0.6}$

Table 4: **Weakly supervised ablation on BFFHQ.** Results are reported on CLIP ResNet-50. "No w" removes confidence weighting, "No warm-up" fixes λ from the beginning of training, and "No $\mathcal{L}_{\text{cmin}}$" removes the minimality regularizer. Bold indicates the best WGA, and underlined denotes the second best WGA.

Method	BFFHQ (1%)		BFFHQ (5%)	
	WGA↑	Avg	WGA↑	Avg
ERM	$63.2_{\pm 2.4}$	$84.7_{\pm 0.5}$	$86.0_{\pm 2.8}$	$94.2_{\pm 0.9}$
Weak bias + Global Adv.	$72.9_{\pm 1.4}$	$91.4_{\pm 0.6}$	$88.2_{\pm 1.4}$	$97.1_{\pm 0.7}$
CMDR-Pseudo w/o w	$73.1_{\pm 1.1}$	$87.8_{\pm 0.6}$	$88.9_{\pm 1.2}$	$97.2_{\pm 0.8}$
CMDR-Pseudo w/o warm-up	$73.7_{\pm 0.8}$	$89.0_{\pm 0.4}$	$89.1_{\pm 1.0}$	$95.3_{\pm 0.6}$
CMDR-Pseudo w/o $\mathcal{L}_{\text{cmin}}$	$\underline{74.0_{\pm 1.3}}$	$89.7_{\pm 0.4}$	$\underline{89.6_{\pm 1.7}}$	$97.8_{\pm 1.5}$
CMDR-Pseudo	$\mathbf{76.0}_{\pm 1.2}$	$88.0_{\pm 0.4}$	$\mathbf{90.0}_{\pm 0.8}$	$95.1_{\pm 1.3}$

over ERM, but remains below full CMDR-Pseudo, highlighting the need for a *conditional* debiasing objective. Removing confidence weighting w or the warm-up of λ consistently reduces WGA, confirming their role in mitigating noisy pseudo labels. Removing $\mathcal{L}_{\text{cmin}}$ slightly increases Avg but lowers WGA, indicating that minimality regularization favors hardest-group robustness over average accuracy.

5 Conclusion

We propose **Conditionally Minimal Sufficient Debiasing Representation (CMDR)** to mitigate spurious correlations and subgroup performance degradation in visual learning. Our analysis establishes a feasibility separation between global invariance and conditional minimal sufficiency under a realizable setting, motivating target-conditioned debiasing while limiting input-mediated nuisances. CMDR combines target-conditioned adversarial debiasing with lightweight variational conditional minimality regularization and supports both supervised debiasing with explicit bias labels and weakly supervised debiasing without training bias labels. Experiments on Waterbirds, CelebA, BAR, MetaShift, and BFFHQ demonstrate strong worst-group robustness across diverse bias types, model architectures, and bias-conflicting ratios. Ablations further validate the advantage of target-conditioned adversarial debiasing over global adversarial debiasing, as well as the contributions of variational conditional minimality regularization and weakly supervised bias discovery. These results suggest that conditional minimal sufficiency provides a principled and practical alternative to global invariance for robust visual representation learning when target and bias variables are statistically dependent. The current evaluation focuses on image classification, while estimating reliable pseudo-bias signals may become more challenging under complex or continuous nuisance factors. Future work will extend CMDR to broader visual and multimodal tasks and improve weakly supervised bias discovery.

AI USE STATEMENT

In this work, we used generative AI tools, specifically ChatGPT (OpenAI), to assist with theoretical derivations and the development and verification of mathematical proofs, as well as for research ideation in the form of literature exploration, summarization, and identification of relevant prior work. We have not used generative AI tools to design or provide feedback on the proposed methodology or experiments, implement the proposed method, process experimental data, or interpret experimental results, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to aid and polish writing, including language editing, improving clarity and readability, and assisting with the organization and presentation of the manuscript. We have reviewed all AI-assisted work; AI-assisted derivations and proofs were independently checked by the authors for mathematical correctness, literature-related suggestions were manually verified against the original sources, and AI-assisted text was reviewed and revised by the authors before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, proofs, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the mitigation of spurious correlations and subgroup performance disparities in visual learning. Our experiments use established benchmark datasets and do not involve new human-subject recruitment or data collection. Some datasets include demographic attributes, such as gender and age, which are used only for bias supervision or subgroup evaluation. These attributes should not be interpreted as complete representations of demographic identity or fairness. Although CMDR is designed to improve subgroup robustness, performance on benchmark groups does not guarantee fairness in real-world settings, particularly for unmeasured or intersecting sources of bias. We therefore recommend careful, domain-specific evaluation before deployment in applications involving sensitive populations or high-stakes decisions.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs of our theoretical results in Appendix A, implementation and training details in Appendix B, and the full pseudo-bias discovery procedure in Appendix C. Additional hyperparameter analyses and experimental details are included in the subsequent appendices to facilitate reproduction of our results.

REFERENCES

Ihab Asaad, Maha Shadaydeh, and Joachim Denzler. Gradient extrapolation for debiased representation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3819–3829. IEEE, 2025.

Sai Siddhartha Chary Aylapuram, Veeraraju Elluru, and Shivang Agarwal. Bias-aware machine unlearning: Towards fairer vision models via controllable forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2571–2579, 2025.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pp. 872–881. PMLR, 2019.

Pu Chen, Linna Wu, and Lei Wang. Ai fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences*, 13(18):10258, 2023.

Sepehr Dehdashtian, Ruozhen He, Yi Li, Guha Balakrishnan, Nuno Vasconcelos, Vicente Ordonez, and Vishnu Naresh Boddeti. Fairness and bias mitigation in computer vision: A survey. *arXiv preprint arXiv:2408.02464*, 2024.

Rajeev Ranjan Dwivedi, Mohammedkaif Mohammedrafiq Kalagond, Niramay Patel, and Vinod Kurmi. Regulating internal alignment flows for robust learning under spurious correlations. In *International Conference on Learning Representations*, volume 2026, pp. 2932–2954, 2026.

Mateo Espinosa Zarlenga, Swami Sankaranarayanan, Jerone TA Andrews, Zohreh Shams, Mateja Jamnik, and Alice Xiang. Efficient bias mitigation without privileged information. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502, 2024.

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

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Madry. Data debiasing with datamodels (d3m): Improving subgroup robustness via data selection. *arXiv preprint arXiv:2406.16846*, 2024.

Taeuk Jang and Xiaoqian Wang. Fades: Fair disentanglement with sensitive relevance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12067–12076, 2024.

Taeuk Jang, Hongchang Gao, Pengyi Shi, and Xiaoqian Wang. Achieving fairness through separability: a unified framework for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 28–36. PMLR, 2024.

Sungho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. In *International Conference on Learning Representations*, volume 2026, pp. 121167–121191, 2026.

Hoin Jung, Junyi Chai, and Xiaoqian Wang. Adversarial latent feature augmentation for fairness. In *The Thirteenth International Conference on Learning Representations*, 2025.

Adarsh Kappiyath, Abhra Chaudhuri, Ajay Kumar Jaiswal, Ziquan Liu, Yunpeng Li, Xiatian Zhu, and Lu Yin. Sebra: Debiasing through self-guided bias ranking. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

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

Jungsoo Lee, Jeonghoon Park, Daeyoung Kim, Juyoung Lee, Edward Choi, and Jaegul Choo. Revisiting the importance of amplifying bias for debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14974–14981, 2023.

Weiwei Li, Junzhuo Liu, Yuanyuan Ren, Yuchen Zheng, Yahao Liu, and Wen Li. Let samples speak: Mitigating spurious correlation by exploiting the clusterness of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15486–15496, 2025.

Weixin Liang, Xinyu Yang, and James Y Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts. In *ICML 2022 Shift Happens Workshop*, 2022.

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

Yujie Lin, Chen Zhao, Minglai Shao, Baoluo Meng, Xujiang Zhao, and Haifeng Chen. Towards counterfactual fairness-aware domain generalization in changing environments. *arXiv preprint arXiv:2309.13005*, 2023.

Yujie Lin, Dong Li, Minglai Shao, Guihong Wan, and Chen Zhao. Fade: Towards fairness-aware generation for domain generalization via classifier-guided score-based diffusion models. *arXiv preprint arXiv:2406.09495*, 2024.

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

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Shenyu Lu, Yipei Wang, and Xiaoqian Wang. Debiasing attention mechanism in transformer without demographics. In *The Twelfth International Conference on Learning Representations*, 2024.

Shenyu Lu, Zhaoying Pan, and Xiaoqian Wang. Think twice: Test-time reasoning for robust clip zero-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2919–2929, 2025.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1295–1305, 2022.

Geon Yeong Park, Chanyong Jung, Sangmin Lee, Jong Chul Ye, and Sang Wan Lee. Self-supervised debiasing using low rank regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12395–12405, 2024.

Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. *ArXiv*, pp. arXiv–2301, 2023.

Maan Qraitem, Kate Saenko, and Bryan A Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20311–20320, 2023.

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

Hamza Riaz and Alan F Smeaton. Resilience of vision transformers for domain generalisation in the presence of out-of-distribution noisy images. *arXiv preprint arXiv:2504.04225*, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16742–16751, 2022.

Agnieszka Słowik and Léon Bottou. On distributionally robust optimization and data rebalancing. In *International Conference on Artificial Intelligence and Statistics*, pp. 1283–1297. PMLR, 2022.

Silpa Vadakkeeveetil Sreelatha, Adarsh Kappiyath, Abhra Chaudhuri, and Anjan Dutta. Denetdm: Debiasing by network depth modulation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 99488–99518. Curran Associates, Inc., 2024. doi: 10.52202/079017-3157. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b40d5797756800c97f3d525c2e4c8357-Paper-Conference.pdf.

Xuebin Tang, Chunchao Li, and Yuanxi Peng. Unsupervised joint adversarial domain adaptation for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

Taro Tezuka and Shizuma Namekawa. Information bottleneck analysis by a conditional mutual information bound. *Entropy*, 23(8):974, 2021.

Bowei Tian, Ruijie Du, and Yanning Shen. Fairvit: Fair vision transformer via adaptive masking. In *European Conference on Computer Vision*, pp. 451–466. Springer, 2024.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8919–8928, 2020.

Chenyu You, Yifei Min, Weicheng Dai, Jasjeet S Sekhon, Lawrence Staib, and James S Duncan. Calibrating multi-modal representations: A pursuit of group robustness without annotations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26140–26150. IEEE, 2024.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhiwen Zhang, Teng Li, Xuebin Tang, Xiang Hu, and Yuanxi Peng. Caevt: Convolutional autoencoder meets lightweight vision transformer for hyperspectral image classification. *Sensors*, 22 (10):3902, 2022.

Beier Zhu, Jiequan Cui, Hanwang Zhang, and Chi Zhang. Project-probe-aggregate: Efficient fine-tuning for group robustness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25487–25496, 2025.

A Theoretical Analysis

We provide a compact theoretical justification for CMDR. Our analysis shows that global bias invariance can be incompatible with target sufficiency, while target-conditioned debiasing can avoid this conflict. We further connect conditional bias dependence to group-risk disparity and conditional minimality to residual nuisance information.

Setup. Let $Y \in \{1, \dots, C\}$ denote the target, Z_b a bias variable, and $H \sim q_\theta(H \mid X)$ the representation induced by the encoder. Since H is generated from X, we have $H \perp (Y, Z_b, U) \mid X$ for any auxiliary nuisance U considered below.

A.1 Global Invariance and Conditional Feasibility

We first characterize why global bias invariance may be overly restrictive.

Proposition 1 (Global invariance–sufficiency trade-off). For any representation H,

$$I(H; Z_b) \geq I(Y; Z_b) - \mathcal{H}(Y \mid H). \tag{12}$$

Hence, if $\mathcal{H}(Y \mid H) < I(Y; Z_b)$, then $I(H; Z_b) > 0$ and exact global bias independence is impossible.

Proof. By the chain rule,

$$\begin{aligned} I(H; Z_b) &= I(Y; Z_b) + I(H; Z_b \mid Y) - I(Y; Z_b \mid H) \\ &\geq I(Y; Z_b) - \mathcal{H}(Y \mid H), \end{aligned} \tag{13}$$

where $I(H; Z_b \mid Y) \geq 0$ and $I(Y; Z_b \mid H) \leq \mathcal{H}(Y \mid H)$. □

This result also yields a simple feasibility separation. Suppose $\mathcal{H}(Y \mid X) = 0$ and $I(Y; Z_b) > 0$. Any target-sufficient representation satisfies $\mathcal{H}(Y \mid H) = 0$, and therefore Eq. 12 implies

$$I(H; Z_b) \geq I(Y; Z_b) > 0.$$

Thus, target sufficiency and exact global invariance cannot be satisfied simultaneously. In contrast, under $\mathcal{H}(Y \mid X) = 0$, Y is a deterministic function of X almost surely. Hence, at the level of the information-theoretic formulation, the ideal representation $H^\star = Y$ satisfies

$$I(Y; X \mid H^\star) = I(H^\star; Z_b \mid Y) = I(H^\star; X \mid Y) = 0, \tag{14}$$

showing that the three CMDR criteria can remain jointly feasible.

The same trade-off can be expressed through prediction error. For a classifier $\hat{Y}$ based on H with $\epsilon = \Pr(\hat{Y} \neq Y)$, Fano's inequality gives

$$I(H; Z_b) \geq I(Y; Z_b) - h_{\mathrm{b}}(\epsilon) - \epsilon \log(C - 1), \tag{15}$$

where h_{b} is the binary entropy. Therefore, enforcing $I(H; Z_b) \leq \delta$ requires

$$h_{\mathrm{b}}(\epsilon) + \epsilon \log(C - 1) \geq I(Y; Z_b) - \delta. \tag{16}$$

Thus, sufficiently strong global invariance can impose an unavoidable prediction-error cost when Y and Z_b are dependent.

A.2 Conditional Dependence and Group Risk

CMDR instead suppresses the residual dependence $I(H; Z_b \mid Y)$.

Proposition 2 (Conditional dependence controls group-risk gaps). Assume discrete groups (Y, Z_b) and

$$\pi_{\min} = \min_{(y,z):p(y,z)>0} p(y, z) > 0.$$

For a bounded loss $\ell(f(H), Y) \in [0, 1]$, define

$$R_{y,z} = \mathbb{E}[\ell(f(H), y) \mid Y = y, Z_b = z], \qquad R_y = \mathbb{E}[\ell(f(H), y) \mid Y = y].$$

Then

$$\max_{(y,z):p(y,z)>0} |R_{y,z} - R_y| \le \sqrt{\frac{I(H; Z_b \mid Y)}{2\pi_{\min}}}. \tag{17}$$

Proof. For each supported (y, z), boundedness of the loss and Pinsker's inequality give

$$|R_{y,z} - R_y| \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(p(H \mid y, z) \| p(H \mid y))}.$$

Moreover,

$$I(H; Z_b \mid Y) = \sum_{y,z} p(y, z) D_{\mathrm{KL}}(p(H \mid y, z) \| p(H \mid y)) .$$

Since all terms are nonnegative,

$$D_{\mathrm{KL}}(p(H \mid y, z) \| p(H \mid y)) \le \frac{I(H; Z_b \mid Y)}{\pi_{\min}},$$

which proves Eq. 17. □

For the 0–1 loss, let $R_{\max} = \max_{(y,z):p(y,z)>0} R_{y,z}$. Then

$$R_{\max} \le \max_{y:p(y)>0} R_y + \sqrt{\frac{I(H; Z_b \mid Y)}{2\pi_{\min}}}, \tag{18}$$

directly connecting conditional bias dependence to worst-group robustness.

A.3 Conditional Minimality

Conditional debiasing only targets the specified Z_b. Other input-mediated nuisances may remain in the representation.

Proposition 3 (Input-mediated nuisance control). For any nuisance U satisfying $U \perp H \mid X, Y$,

$$I(H; U \mid Y) \le I(H; X \mid Y). \tag{19}$$

Proof. Since $I(H; U \mid X, Y) = 0$,

$$I(H; X, U \mid Y) = I(H; X \mid Y), \tag{20}$$
$$I(H; X, U \mid Y) = I(H; U \mid Y) + I(H; X \mid U, Y) \ge I(H; U \mid Y). \tag{21}$$

Combining the two expressions proves the result. □

Hence, minimizing $I(H; X \mid Y)$ limits residual dependence on input-mediated nuisances beyond the explicitly modeled bias. For Z_b itself,

$$I(H; X \mid Y) = I(H; Z_b \mid Y) + I(H; X \mid Y, Z_b), \tag{22}$$

where we use $I(H; Z_b \mid X, Y) = 0$. This shows that conditional debiasing and conditional minimality play complementary roles.

A.4 Tractable Surrogates

For conditional minimality, define

$$q_\theta(H \mid y) = \int q_\theta(H \mid x) p(x \mid y) \, dx.$$

For any variational prior $r_\eta(H \mid Y)$,

$$\begin{aligned} \mathbb{E}_{X,Y} D_{\mathrm{KL}}(q_\theta(H \mid X) \| r_\eta(H \mid Y)) &= I(H; X \mid Y) \\ &\quad + \mathbb{E}_Y D_{\mathrm{KL}}(q_\theta(H \mid Y) \| r_\eta(H \mid Y)) \\ &\ge I(H; X \mid Y). \end{aligned} \tag{23}$$

Thus, $\mathcal{L}_{\text{cmin}}$ provides a variational upper bound on target-conditioned input redundancy.

For target sufficiency,

$$I(Y; X \mid H) = \mathcal{H}(Y \mid H) - \mathcal{H}(Y \mid X), \tag{24}$$

which motivates predictive cross-entropy.

For discrete Z_b, an optimal conditional discriminator satisfies

$$\min_g \mathbb{E}[-\log g(Z_b \mid h, Y)] = \mathcal{H}(Z_b \mid h, Y).$$

Since

$$I(h; Z_b \mid Y) = \mathcal{H}(Z_b \mid Y) - \mathcal{H}(Z_b \mid h, Y),$$

adversarially increasing the optimal discriminator loss reduces $I(h; Z_b \mid Y)$. Because $Z_b \to h \to H$ conditioned on Y,

$$I(H; Z_b \mid Y) \leq I(h; Z_b \mid Y).$$

For continuous pseudo-biases, we use MSE as a practical surrogate for suppressing target-conditioned predictability, without claiming exact conditional mutual-information minimization.

Overall, CMDR is motivated by three complementary observations: global invariance may conflict with target sufficiency, conditional debiasing controls target-conditioned group disparity, and conditional minimality limits residual input-mediated nuisance information.

B Experimental Details

Throughout our experiments, we adopted the same CLIP ResNet-50 and CLIP ViT-L/14 backbones for all compared methods whenever applicable. The corresponding model architectures and pretrained weights were obtained from Hugging Face using `timm/resnet50_clip.openai` and `openai/clip-vit-large-patch14`, respectively. Each dataset was divided into training, validation, and test sets following the standard splits. For group-agnostic methods, subgroup annotations were not used during training or pseudo-bias discovery, and were only used for validation model selection and test-time worst-group accuracy evaluation. For subgroup-supervised methods, the corresponding bias or subgroup annotations were used according to the original method design.

For most baseline methods, we used a batch size of 128, a learning rate of 1×10^{-4}, the cross-entropy loss, and the Adam optimizer. The number of training epochs was selected according to the dataset: 50 for Waterbirds, 20 for CelebA, 200 for BAR, 50 for MetaShift, and 40 for BFFHQ. We report worst-group accuracy as the primary metric and average accuracy as the secondary metric.

For CMDR, the pretrained backbone was followed by a lightweight projection layer and a linear classification head. In CMDR-Pseudo, pseudo-bias vectors and confidence weights were first estimated in a frozen pretrained feature space and then used as soft supervision for the target-conditioned bias discriminator. In CMDR-Bias, we used the ground-truth bias labels and set the confidence weight to 1 for all samples. The conditional bias discriminator was trained with a gradient reversal layer. For discrete bias labels, the bias prediction loss was implemented as cross-entropy; for continuous pseudo-bias vectors, it was implemented as mean squared error. Unless otherwise specified, we set the conditional minimality coefficient to $\beta = 1 \times 10^{-4}$ and used a warm-up schedule for the conditional debiasing coefficient λ. For the variational conditional minimality loss, we used learnable class-specific means m_y and diagonal log-variances to parameterize $r_\eta(H \mid y)$. The diagonal variances were parameterized by log-variances to ensure positivity, and τ^2 was fixed throughout training. All reported results are averaged over three independent random seeds. Ablation results are reported as mean $\pm$ standard deviation over three seeds.

C Weakly Supervised Pseudo-Bias Discovery

This section describes how CMDR-Pseudo constructs pseudo-bias supervision when group annotations are unavailable. Given a training set $\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^N$, our goal is to assign each sample a continuous pseudo-bias vector z_i^b and a confidence weight $w_i \in [0, 1]$. The pseudo-bias is not designed to recover the ground-truth bias attribute. Instead, it provides a soft nuisance direction that captures visual factors correlated with background, style, and augmentation-sensitive appearance cues. No group annotations are used during this discovery stage.

Weak visual signals. We first extract a frozen global representation

$$h_i = E(x_i), \tag{25}$$

where E is a pretrained visual encoder and is kept fixed throughout pseudo-bias discovery. For each image, we also compute a weak visual descriptor s_i by concatenating three types of cues:

$$s_i = [s_i^{\mathrm{bg}};\ s_i^{\mathrm{style}};\ s_i^{\mathrm{aug}}]. \tag{26}$$

Here, s_i^{bg} contains coarse border color statistics, luminance, and contrast, which are useful for capturing background shortcuts. The term s_i^{style} is obtained from Gram statistics of intermediate VGG features and describes low-level texture and style. Finally, s_i^{aug} measures the absolute feature difference between two independently sampled strong augmentations of the same image:

$$s_i^{\mathrm{aug}} = \left| E(a_1(x_i)) - E(a_2(x_i)) \right|, \tag{27}$$

where a_1 and a_2 are sampled from a strong augmentation pipeline including random resized cropping, color jittering, grayscale conversion, horizontal flipping, and Gaussian blur.

PLS projection. After collecting all samples, we form the feature matrix $\mathbf{H} = [h_1, \ldots, h_N]^\top$ and the weak-signal matrix $\mathbf{S} = [s_1, \ldots, s_N]^\top$. Both matrices are standardized feature-wise. We then fit a partial least squares regression model from $\widetilde{\mathbf{H}}$ to $\widetilde{\mathbf{S}}$:

$$\mathrm{PLS} : \widetilde{\mathbf{H}} \to \widetilde{\mathbf{S}}, \tag{28}$$

where $\widetilde{\mathbf{H}}$ and $\widetilde{\mathbf{S}}$ denote standardized matrices. The latent PLS variables are used as the pseudo-bias representation:

$$Z = \mathrm{PLS.transform}(\widetilde{\mathbf{H}}) \in \mathbb{R}^{N \times d_b}. \tag{29}$$

We use $d_b = 2$ in the main experiments. The resulting matrix Z is standardized again, and its i-th row is used as z_i^b. This procedure keeps low-dimensional directions in the frozen representation space that are most predictive of the weak nuisance descriptors.

Confidence weighting. Since the discovered pseudo-bias is estimated without group labels, we down-weight unreliable samples. The confidence score combines three complementary criteria. First, we measure how well the fitted PLS model reconstructs the weak signal; samples with smaller reconstruction error receive higher confidence. Second, we measure the magnitude of the standardized pseudo-bias vector, since very weak latent responses are less informative. Third, we estimate bootstrap stability by refitting PLS on resampled training sets and measuring the variance of the induced pseudo-bias vectors after component-wise sign alignment. The final confidence weight is

$$w_i = \mathrm{clip}\left(0.4 w_i^{\mathrm{rec}} + 0.3 w_i^{\mathrm{mag}} + 0.3 w_i^{\mathrm{stab}}, 0, 1\right), \tag{30}$$

where w_i^{rec}, w_i^{mag}, and w_i^{stab} denote the reconstruction, magnitude, and stability scores, respectively. In our implementation, bootstrap stability is estimated with five resampling rounds.

Use in CMDR training. The discovered tuple $\{(z_i^b, w_i)\}_{i=1}^N$ is cached and loaded by the training dataset according to the original sample index. During CMDR-Pseudo training, z_i^b serves as the target of the target-conditioned bias discriminator, while w_i weights the conditional debiasing loss:

$$\mathcal{L}_{\mathrm{cind}} = \frac{1}{N} \sum_{i=1}^{N} w_i \ell_b \left(g_\omega(\mathrm{GRL}(\phi_\theta(x_i)), y_i), z_i^b \right). \tag{31}$$

For continuous pseudo-bias vectors, ℓ_b is implemented as mean squared error. When true bias labels are available, we replace z_i^b with the ground-truth bias label and set $w_i = 1$.

D HYPERPARAMETER SENSITIVITY

Table 5 studies the effects of λ and β. CMDR is stable across a relatively wide range of values. Increasing λ from 0.1 to 0.5 improves WGA, while performance remains comparable at $\lambda = 0.7$ and slightly degrades at $\lambda = 1.0$, suggesting that overly strong adversarial debiasing may suppress task semantics. For β, a small conditional minimality regularizer ($\beta = 1 \times 10^{-4}$) improves WGA over $\beta = 0$, whereas larger values cause mild degradation due to an overly strong information penalty. Overall, CMDR performs best with moderate conditional debiasing and lightweight conditional minimality regularization.

Algorithm 1 Weakly Supervised Pseudo-Bias Discovery

Require: Training set $\mathcal{D}_{\mathrm{tr}}$ without group annotations; frozen encoder E; weak descriptor extractor; PLS dimension d_b; bootstrap rounds B.
Ensure: Pseudo-bias vectors $\{z_i^b\}_{i=1}^N$ and confidence weights $\{w_i\}_{i=1}^N$.
1: **for** each training sample x_i **do**
2: Extract frozen global feature $h_i \leftarrow E(x_i)$.
3: Build weak descriptor $s_i = [s_i^{\mathrm{bg}}; s_i^{\mathrm{style}}; s_i^{\mathrm{aug}}]$.
4: **end for**
5: Standardize feature matrix $\mathbf{H}$ and weak-signal matrix $\mathbf{S}$.
6: Fit PLS regression from $\widetilde{\mathbf{H}}$ to $\widetilde{\mathbf{S}}$.
7: Obtain latent pseudo-bias matrix $Z \leftarrow \mathrm{PLS.transform}(\widetilde{\mathbf{H}})$.
8: Standardize Z and set the i-th row as z_i^b.
9: Compute confidence scores from reconstruction quality, pseudo-bias magnitude, and bootstrap stability.
10: Combine the scores to obtain w_i.
11: Cache $\{(z_i^b, w_i)\}_{i=1}^N$ for CMDR-Pseudo training.

Table 5: **Sensitivity to λ and β on Waterbirds.** Results are reported on CLIP ViT-L/14. Bold indicates the best WGA, and underlined denotes the second best WGA.

λ	β	WGA↑	Avg
0.1	1×10^{-4}	89.242	95.114
0.3	1×10^{-4}	90.181	94.882
0.5	1×10^{-4}	**90.810**	95.560
0.7	1×10^{-4}	<u>90.687</u>	94.411
1.0	1×10^{-4}	90.356	94.027
0.5	0	88.864	94.991
0.5	5×10^{-4}	90.525	94.219

E Necessity and Effectiveness of Conditional Minimality

The conditional adversarial objective provides the main debiasing signal by reducing $I(H; Z^b \mid Y)$. However, it does not prevent H from retaining other within class nuisance or sample specific information. Conditional minimality addresses this limitation by reducing $I(H; X \mid Y)$, thereby compressing residual input information, while the classification objective preserves target semantics.

Under explicit bias supervision, adding conditional minimality consistently improves WGA without reducing Avg:

Dataset	Cond. Adv.	+ Cond. Min.	ΔWGA
Waterbirds	88.9 / 95.0	**90.8 / 95.6**	+1.9
CelebA	87.3 / 90.8	**88.8 / 91.1**	+1.5
BAR	80.6 / 96.5	**84.5 / 97.3**	+3.9

Each entry reports WGA / Avg.

Its effect is less consistent under weak supervision. On BFFHQ at the 1% conflict ratio, it improves WGA by 2.0 points while Avg decreases by 1.7 points. At the 5% conflict ratio, WGA improves by 0.4 points while Avg decreases by 2.7 points. This indicates that conditional minimality primarily targets robustness and may introduce a tradeoff in the presence of imperfect pseudo bias signals.

To isolate its practical role, we conduct a controlled nuisance experiment on Waterbirds. Each training image is assigned an independently sampled low frequency Gaussian perturbation that remains fixed across epochs. The perturbation is sampled independently of the target and bias labels from the same distribution for all classes and groups. At test time, a new perturbation is sampled for each image. Checkpoints are selected using clean validation WGA.

Method	Noise std	Clean WGA	Resampled WGA	Drop	Drift
Cond. Adv.	0.1	62.129	60.710	1.419	0.208
Full CMDR	0.1	**63.060**	**62.439**	**0.621**	**0.182**
Cond. Adv.	0.2	64.174	60.125	4.049	0.245
Full CMDR	0.2	**64.568**	**62.111**	**2.457**	**0.213**

Here, Drift measures the representation change between the clean and resampled versions of the same image. Full CMDR retains similar clean performance while producing smaller WGA drops and representation changes. These results show that conditional minimality reduces sensitivity to residual within class nuisance that is not addressed by the specified bias adversary. It is not required for the basic conditional adversarial effect, but it provides additional robustness and enforces the minimality principle of CMDR.

F Robustness without Bias-Conflicting Training Samples

CMDR-Bias is not applicable to the 0% conflict setting. When the observed bias is deterministic given the target label, no within-class bias variation is available to provide a meaningful conditional adversarial signal.

CMDR-Pseudo remains applicable because it requires neither bias annotations nor explicitly identified conflicting samples. Instead, it discovers continuous nuisance directions from a frozen pretrained representation. These directions need not correspond exactly to a human-defined bias attribute, but may capture latent nuisance variations encoded in the pretrained feature space. We do not claim guaranteed recovery when no shortcut-related variation is detectable.

We evaluate this setting on Waterbirds and MetaShift (Cat vs. Dog) using CLIP ResNet-50. For the 0% setting, all bias-conflicting samples are removed from training while the validation and test sets remain unchanged. We report all four subgroup accuracies together with WGA and Avg.

Dataset	Conflict	Method	Group0	Group1	Group2	Group3	WGA	Avg
Waterbirds	0%	ERM	99.3	30.7	12.8	93.5	12.8	94.4
		CMDR-Pseudo	95.5	37.6	39.6	93.0	37.6 (+24.8)	92.1
	5%	ERM	99.5	77.2	43.8	91.0	43.8	96.1
		CMDR-Pseudo	95.9	75.7	70.4	89.9	70.4 (+26.6)	93.5
MetaShift	0%	ERM	98.8	70.2	42.3	99.2	42.3	98.6
		CMDR-Pseudo	99.2	68.7	76.9	99.6	68.7 (+26.4)	99.1
	1%	ERM	98.9	77.3	80.3	99.8	77.3	99.2
		CMDR-Pseudo	99.4	85.0	82.5	99.7	82.5 (+5.2)	99.4
	5%	ERM	98.8	90.2	93.7	99.7	90.2	98.9
		CMDR-Pseudo	98.9	93.5	93.5	99.6	93.5 (+3.3)	99.0

At 0% conflict, CMDR-Pseudo improves WGA from 12.8 to 37.6 on Waterbirds and from 42.3 to 68.7 on MetaShift, corresponding to gains of 24.8 and 26.4 points, respectively. The gains primarily come from substantially improved robustness on the previously unseen conflict groups. Avg is preserved on MetaShift (98.6 to 99.1), while Waterbirds exhibits a moderate robustness–accuracy trade-off (94.4 to 92.1). CMDR-Pseudo also consistently improves WGA under nonzero conflict ratios.

These results show that pretrained latent directions provide a practical fallback when explicit conflict evidence is absent. Nevertheless, this ability depends on shortcut-related variation being detectable in the pretrained representation and does not remove the fundamental identifiability limitation.

G Use of Validation Group Labels for Model Selection

Validation group labels are used only for checkpoint selection and never enter pseudo-bias discovery or gradient updates. In real-world datasets, bias annotations are often unavailable and costly to

obtain, while training distributions are typically highly imbalanced. Reusing the validation set for supervised debiasing would change the supervision setting and contaminate the held-out split. For example, Waterbirds has train group counts $[3498, 56, 184, 1057]$, whereas its validation and test distributions, $[467, 133, 466, 133]$ and $[2255, 642, 2255, 642]$, are nearly identical. Fine-tuning on the 1,199 validation samples would therefore inject test-like group supervision. Since such validation–test alignment is generally unknown, we keep validation data strictly out of training. Validation WGA is used only as a simple and consistent model-selection criterion aligned with the primary robustness metric.

H Evaluation of Pseudo-Bias Signal Quality

To evaluate the quality of the discovered pseudo-bias signals beyond the CelebA example shown in the main paper, we report precision, recall, and AUROC for every target class on Waterbirds, CelebA, BAR, MetaShift, and BFFHQ. Overall, the pseudo-bias signals provide informative separation for most target classes, although their quality varies across datasets, particularly under highly imbalanced bias-group distributions. CelebA shows the strongest and most consistent results, while Waterbirds, MetaShift, BFFHQ, and different BAR classes exhibit varying levels of pseudo-bias quality.

The main failure case is BAR class 2, for which precision and recall are both zero. The bias-aligned and bias-conflicting sample counts across the six BAR classes are $[296, 490, 133, 249, 306, 287]$ and $[14, 24, 6, 12, 15, 14]$, respectively. Class 2 contains only six bias-conflicting samples, and the pseudo-bias estimator fails to select any of them. This failure coincides with the extremely limited number of minority samples, highlighting the difficulty of pseudo-bias discovery under severe group imbalance.

Dataset	Target class	Precision	Recall	AUROC
Waterbirds	0	0.341	0.492	0.818 [0.751, 0.881]
	1	0.231	0.396	0.747 [0.674, 0.812]
CelebA	0	0.280	0.781	0.962 [0.947, 0.974]
	1	0.585	0.899	0.956 [0.918, 0.988]
BAR	0	0.108	0.236	0.774 [0.691, 0.844]
	1	0.558	0.585	0.969 [0.952, 0.980]
	2	0	0	0.643 [0.508, 0.805]
	3	0.171	0.432	0.767 [0.622, 0.887]
	4	0.074	0.400	0.739 [0.534, 0.985]
	5	0.417	0.909	0.991 [0.973, 1.000]
MetaShift (5%)	0	0.162	0.206	0.749 [0.699, 0.803]
	1	0.103	0.122	0.599 [0.561, 0.635]
BFFHQ (5%)	0	0.256	0.303	0.701 [0.653, 0.857]
	1	0.189	0.213	0.662 [0.615, 0.704]

Additional Visualization of Pseudo-Bias Signals. To complement the visualization analysis of pseudo-bias quality in the main paper, we provide t-SNE visualizations for all target classes on Waterbirds and CelebA. These additional results provide a more complete view of how the discovered pseudo-bias signals separate bias-conflicting and bias-aligned samples across different target classes.

I Pseudo-Bias Extractor Independence and Backbone Choice

All pseudo-bias labels used in the main experiments are generated offline by applying the weakly supervised PLS-based procedure to features extracted from a frozen CLIP ViT-L/14 model. We use a large-scale pretrained model because its features are semantically rich and broadly transferable, which is consistent with our goal of discovering latent bias-related variations beyond those learned from the biased downstream dataset.

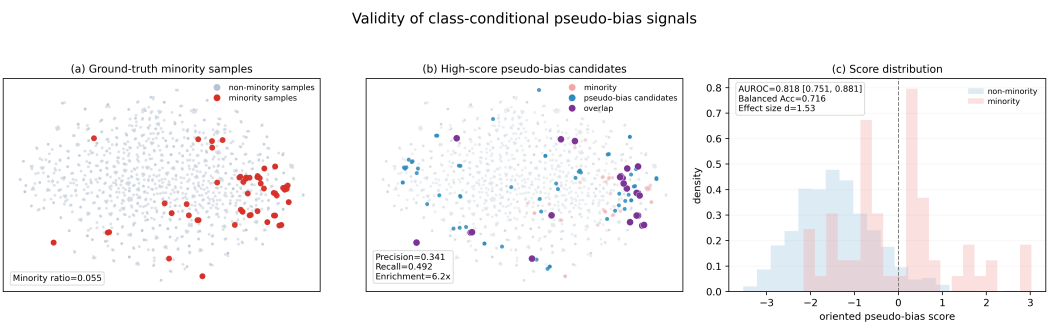

Figure 3: t-SNE visualization of pseudo-bias signals on Waterbirds for target class 0.

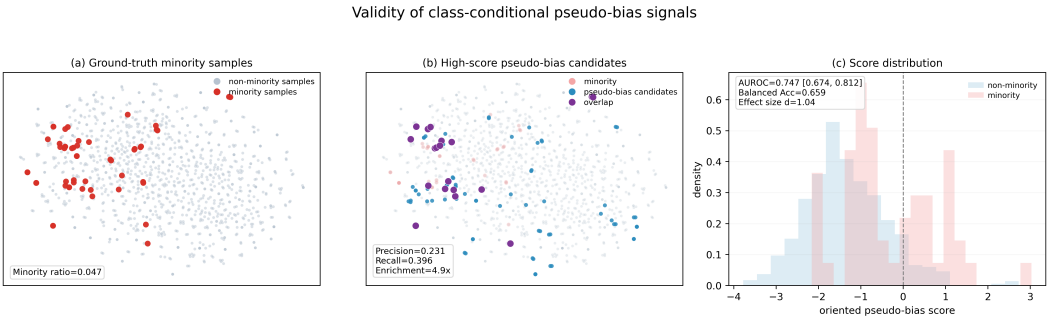

Figure 4: t-SNE visualization of pseudo-bias signals on Waterbirds for target class 1.

The pseudo-bias extractor and the debiased model are not jointly optimized. Pseudo-bias labels are generated before CMDR training and remain fixed throughout training, and no gradient from CMDR is propagated to the frozen extractor. Moreover, CMDR does not directly remove the PLS directions from the CLIP ViT-L/14 feature space; the resulting pseudo-labels are used only as weak supervision for the conditional adversarial objective. Therefore, the discovery and debiasing stages do not form a jointly optimized feedback loop.

For the CLIP ResNet-50 results in the main experiments, the pseudo-bias extractor and the debiased model use different backbone architectures. To further examine the effect of extractor backbone choice, we fix the debiased backbone as CLIP ResNet-50 and replace the original CLIP ViT-L/14 extractor with a frozen CLIP ResNet-50, while keeping the PLS procedure and all CMDR settings unchanged.

Pseudo-bias extractor	Debiased backbone	Waterbirds		CelebA	
		WGA	Avg	WGA	Avg
CLIP ViT-L/14	CLIP ResNet-50	70.405	93.483	77.778	96.035
CLIP ResNet-50	CLIP ResNet-50	71.028	91.949	75.786	95.437

We also compare the quality of the pseudo-bias labels produced by the two frozen extractors:

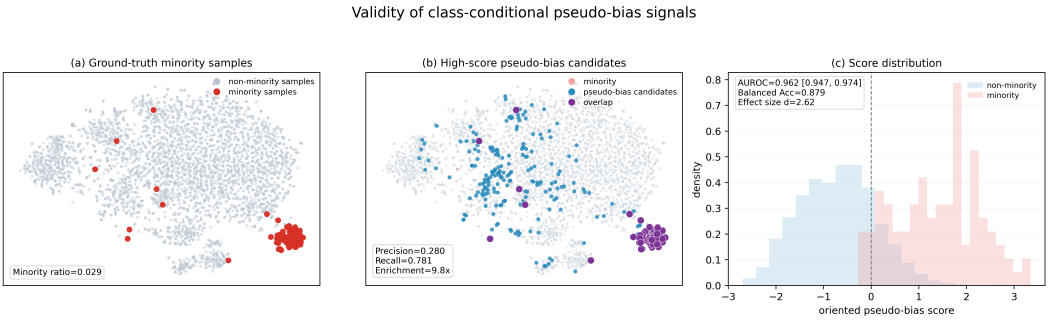

Figure 5: t-SNE visualization of pseudo-bias signals on CelebA for target class 0.

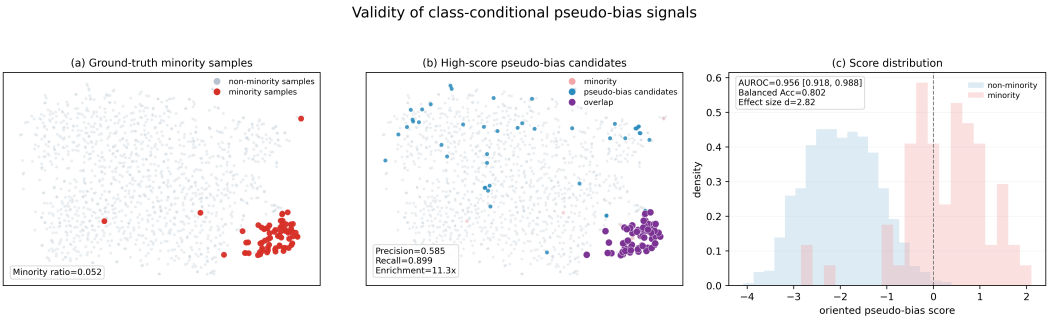

Figure 6: t-SNE visualization of pseudo-bias signals on CelebA for target class 1.

Extractor	Dataset	Target class	Precision	Recall	AUROC
CLIP ViT-L/14	Waterbirds	0	0.341	0.492	0.818
		1	0.231	0.396	0.747
	CelebA	0	0.280	0.781	0.962
		1	0.585	0.899	0.956
CLIP ResNet-50	Waterbirds	0	0.332	0.437	0.798
		1	0.238	0.392	0.756
	CelebA	0	0.235	0.607	0.890
		1	0.454	0.595	0.926

J Comparison with Additional Recent Baselines

We further compare CMDR with three recent debiasing methods, including DeNetDM (Sreelatha et al., 2024), GERNE (Asaad et al., 2025), and Sebra (Kappiyath et al., 2025), on Waterbirds and CelebA using both CLIP ResNet-50 and CLIP ViT-L/14. For a fair comparison with CMDR-Pseudo, we use the unknown-attribute variant of GERNE, which does not require ground-truth bias annotations. DeNetDM, GERNE (unknown attribute), Sebra, and CMDR-Pseudo are compared without bias annotations, while GERNE (known attribute) and CMDR-Bias are evaluated under bias supervision. All methods use the same dataset splits and evaluation metrics.

Method	Bias labels	CLIP ResNet-50				CLIP ViT-L/14			
		Waterbirds		CelebA		Waterbirds		CelebA	
		WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg
DeNetDM	No	50.554	52.390	50.000	56.058	54.590	96.613	77.222	92.966
GERNE (unknown attribute)	No	25.545	80.822	56.542	95.835	81.931	97.494	83.185	85.216
Sebra	No	60.125	94.412	46.667	97.776	70.717	97.635	39.444	97.645
CMDR-Pseudo	No	70.405	93.483	77.778	96.035	83.178	97.777	78.667	93.878
GERNE (known attribute)	Yes	66.044	90.528	87.893	94.848	77.222	85.423	88.215	92.213
CMDR-Bias	Yes	74.455	94.134	88.045	93.152	90.810	95.560	88.816	91.070

The additional results provide direct comparisons with recent architecture-based, optimization-based, and self-guided debiasing methods under matched datasets, backbones, supervision settings, and evaluation metrics.

K Effect of conditional minimality on WGA and Avg

BFFHQ at the 5% conflict ratio exhibits a trade-off between robustness and average accuracy. Adding conditional minimality improves WGA from 89.6 to 90.0, while Avg decreases from 97.8 to 95.1. Since WGA is determined by the weakest subgroup, this improvement can coexist with reduced accuracy on the larger and stronger groups.

To examine whether the decrease results from excessive compression, we additionally vary the conditional minimality weight only on BFFHQ at the 5% conflict ratio:

Minimality weight	WGA	Avg
0	89.6	97.8
$0.5\beta_{\text{cmin}}$	90.8	95.3
β_{cmin}	90.0	95.1

The results indicate that conditional minimality controls the balance between WGA and Avg. Under finite data and imperfect pseudo-bias signals, a large weight may suppress some useful within-class variation, while a moderate weight can reduce excessive compression.

L Reliance on alternative shortcuts

Suppressing one shortcut does not guarantee that the model will not exploit another unmeasured shortcut. To examine this possibility, we conduct a post hoc analysis on CelebA using four additional attributes that are not used by CMDR.

CMDR-Bias is trained on CelebA with blond hair as the target attribute and male as the bias attribute. For each additional attribute, we partition the validation set according to the blond hair label and the attribute value, and report the minimum target accuracy across all nonempty subgroups.

Method	Male	Black Hair	Heavy Makeup	Smiling	Straight Hair
ERM	46.67	82.62	80.30	86.32	87.28
CMDR-Bias	88.05	83.75	87.27	87.72	88.64

CMDR substantially improves robustness across the intended gender groups and does not reduce the minimum subgroup accuracy under any of the four additional attributes. This provides evidence that the improvement is not accompanied by an obvious shift toward these measured alternative shortcuts. We nevertheless acknowledge that unannotated shortcuts cannot be completely excluded.