WHEN DOES NOISE SENSITIVITY REVEAL HEAD-ACCESSIBLE SHORTCUTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Can mild input noise reveal which coordinates a frozen vision model uses as shortcuts? We measure the class-conditional change of each pooled representation coordinate and use this score in Noise-based Class-Conditional Sensitivity Pruning (NCSP). The method masks sensitive connections in each row of a linear classifier and refits the unmasked head parameters on clean labeled data. The score directly measures brittleness; interpreting it as a shortcut signal additionally requires nuisance-linked coordinates to be head-accessible, coordinate-aligned, and more probe-sensitive than stable semantic alternatives. Under sub-exponential tails, empirical top-k masks nearly minimize retained sensitivity mass; the constrained refit is convex and norm-controlled; and cross-row conditions yield prediction and KL-stability bounds. The full pipeline raises BAR accuracy from 60.51% to 89.64%, Waterbirds WGA from 63.74% to 79.43%, MetaShift WGA from 69.60% to 86.05%, and FB-CMNIST accuracy at $q = 0.99$ from 25.5% to 91.9%. Matched controls qualify these gains: clean-data head refitting accounts for a large share of the improvement, sensitivity masks do not uniformly beat random or magnitude-based masks, and a 20-class stress test shows no sensitivity-specific advantage. NCSP is therefore a conditional audit and intervention for a particular representation and probe, not a general guarantee of worst-group performance.

1 INTRODUCTION

Modern visual classifiers can achieve high average accuracy while relying on background, texture, color, or context cues that fail under subpopulation and context shifts (Lapuschkin et al., 2019; Geirhos et al., 2020; Wang et al., 2020; Beery et al., 2018). Many methods address this problem by retraining with group annotations, environments, or inferred hard examples (Sagawa et al., 2019; Koh et al., 2021; Arjovsky et al., 2019; Krueger et al., 2021; Nam et al., 2020; Liu et al., 2021; Creager et al., 2021; Sohoni et al., 2020; Eyuboglu et al., 2022). We study a narrower deployment setting: the model is already trained, the backbone will remain frozen, and nuisance labels may be unavailable.

Our question is whether the representation–head interface contains a usable signal for auditing this model. We measure the class-conditional response of each pooled representation coordinate to mild input noise. This quantity is a *brittleness diagnostic*; it identifies a shortcut only when nuisance-linked coordinates are more probe-sensitive than stable semantic alternatives and are accessible through the final linear head. We call this condition a *stability–semantics gap*. It does not require every shortcut to be high-frequency, and it does not follow from sensitivity alone.

Noise-based Class-Conditional Sensitivity Pruning (NCSP) uses this score to mask the most sensitive connections in each class row of a linear head, then refits the unmasked head parameters on clean labeled data. The row-specific mask preserves coordinates that may be brittle for one class but useful for another. The refit is a constrained multiclass logistic-regression problem; the backbone never changes.

The method is conditional and coordinate-dependent. Controlled comparisons show that clean-data head refitting explains much of the improvement over ERM, and sensitivity masks do not uniformly beat random or magnitude-based masks. Label-only validation can screen out an unsupported intervention, but it cannot guarantee worst-group model selection. We therefore evaluate both positive and negative cases and retain the original model when the diagnostic premise is not supported.

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.

Contributions.

1. We introduce a class-conditional finite-noise brittleness diagnostic for frozen representations and state the additional conditions required to interpret it as evidence of a head-accessible shortcut.
2. We study a row-specific masking and constrained-refit pipeline that uses target labels but no nuisance labels during intervention.
3. We prove finite-sample sensitivity-ranking guarantees under sub-exponential tails, convexity and norm control of the refit, and row-matched and cross-row stability bounds.
4. We evaluate the method and matched controls on BAR, Waterbirds, MetaShift, FB-CMNIST, and a 20-class NICO++ stress test, reporting where sensitivity ranking helps and where it does not.

2 Related work

Spurious correlations and group robustness. Shortcut learning has been documented across natural and synthetic settings, where texture, color, background, or context can dominate decision rules despite being non-causal for the target label (Geirhos et al., 2020; Wang et al., 2020; Ye et al., 2024). Group-robust objectives directly optimize worst-case subpopulation performance when group labels are available (Sagawa et al., 2019; Koh et al., 2021). Invariant methods such as IRM and REx seek predictors that remain stable across environments (Arjovsky et al., 2019; Krueger et al., 2021). NCSP targets a complementary setting: a biased model is already trained, group labels are not used for debiasing or model selection, and only the final linear head is modified.

Label-free bias discovery. Learning from Failure (LfF) (Nam et al., 2020), Just Train Twice (JTT) (Liu et al., 2021), Environment Inference for Invariant Learning (EIIL) (Creager et al., 2021), No Subclass Left Behind (NSLB) (Sohoni et al., 2020), Domino (Eyuboglu et al., 2022), Spotlight (d'Eon et al., 2022), Frustratingly Easy Environment Discovery (FEED) (Zare & Nguyen, 2024), Cross-Risk Minimization (XRM) (Pezeshki et al., 2024), and anomaly-detection-based debiasing approaches (Pastore et al., 2025) reduce dependence on group annotations by constructing proxy groups, hard examples, hidden slices, environments, or anomalous subpopulations from training dynamics, losses, representations, or model errors. Recent concept- and vision-language-based approaches instead expose spurious attributes using external semantic information (Zhang et al., 2024; Dammu & Shah, 2023). These methods operate primarily at the sample, environment, slice, or concept level and often require retraining or an auxiliary discovery stage. NCSP instead audits a trained representation at the coordinate level and modifies only class-specific linear-head connections.

Feature-level and last-layer interventions. Feature-level debiasing modifies representations, activations, or classifier heads to reduce shortcut reliance. Learning not to Learn and ReBias use biased representations or adversarial objectives (Kim et al., 2019; Bahng et al., 2020); SiFER suppresses easily learned features (Shah et al., 2020; Tiwari & Shenoy, 2023); and DFR shows that last-layer retraining can be highly effective when group-balanced validation data are available (Kirichenko et al., 2023). Recent activation-erasing and representation-regularization methods also intervene after representation learning (He et al., 2025; Wen et al., 2025). NCSP is closest to last-layer methods, but its mask is discovered from perturbation sensitivity and its refit uses only label-supervised clean data.

Noise sensitivity, non-robust features, and pruning. Perturbations are widely used for adversarial training, corruption robustness, augmentation, and certification (Madry et al., 2018; Hendrycks & Dietterich, 2019; Hendrycks et al., 2020; Cohen et al., 2019). Related work shows that standard models use non-robust yet predictive features (Ilyas et al., 2019), and that high-frequency or surface-statistical cues can shape generalization (Jo & Bengio, 2017; Yin et al., 2019; Wang et al., 2020). NCSP uses perturbations in the reverse direction, to diagnose unstable representation coordinates rather than to train global robustness. This distinguishes it from compression pruning (LeCun et al., 1989; Han et al., 2016; Frankle & Carbin, 2019), which can amplify bias when shortcut features acquire large weights (Hooker et al., 2020).

3 PROBLEM SETUP AND METHOD

3.1 SETUP

Let $(X, Y, Z) \sim P_{\mathrm{tr}}$, where $X \in \mathcal{X}$ is an input, $Y \in \{1, \ldots, C\}$ is the target label, and Z is a nuisance attribute. Spurious correlation means that $P_{\mathrm{tr}}(Z \mid Y)$ is skewed, while robust prediction should remain accurate under shifts in $P(Z \mid Y)$ that preserve the semantics of Y. NCSP does not use Z during sensitivity estimation, pruning, refitting, or model selection; when group labels are available, they are used only for evaluation.

We assume a trained feature extractor $\phi : \mathcal{X} \to \mathbb{R}^d$ and a linear head $W \in \mathbb{R}^{C \times d}$. The class-$c$ logit is

$$z_c(x) = \langle W_c, \phi(x) \rangle, \tag{1}$$

where W_c is the c-th row of W. The goal is to reduce reliance on nuisance-linked coordinates of $\phi(x)$ while preserving enough label-semantic coordinates for accurate classification.

3.2 CLASS-CONDITIONAL NOISE SENSITIVITY

For a Gaussian probe $T_\sigma(x) = x + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, define the population class-conditional sensitivity of coordinate j as

$$S_{c,j} = \mathbb{E}[|\phi_j(T_\sigma(X)) - \phi_j(X)| \mid Y = c]\,. \tag{2}$$

Given N_e examples $\mathcal{S}_c = \{x_i : y_i = c\}$, NCSP estimates

$$d_{c,j} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} |\phi_j(T_\sigma(x)) - \phi_j(x)|\,. \tag{3}$$

The probe is a local stress test and does not attempt to reproduce the test distribution. Coordinates that respond strongly to a mild label-preserving perturbation are treated as candidates for shortcut-linked brittleness.

3.3 SENSITIVITY PRUNING AND CONSTRAINED HEAD REFIT

For pruning ratio $\tau \in (0, 1)$ and $k = \lceil \tau d \rceil$, NCSP forms a class-specific index set

$$P_c = \mathrm{TopK}(d_c, k), \tag{4}$$

where $\mathrm{TopK}(d_c, k)$ returns the indices of the k largest entries of d_c. It then constrains $W_{c,j} = 0$ for all $j \in P_c$ and re-optimizes the remaining head parameters:

$$\begin{aligned} \min_{W \in \mathbb{R}^{C \times d}} \quad & \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{ce}}(W\phi(x_i), y_i) + \frac{\lambda}{2} \|W\|_F^2 \\ \text{s.t.} \quad & W_{c,j} = 0 \qquad \forall c \in [C],\ j \in P_c. \end{aligned} \tag{5}$$

Class conditioning matters because the same coordinate may be brittle for one class but useful for another, especially when different labels co-occur with different contexts. Row-specific masks therefore avoid the overly blunt intervention of globally deleting the same feature dimensions for all classes. We sort by decreasing sensitivity and use coordinate index as a deterministic secondary key. Tie-breaking makes the implementation reproducible but does not supply a sensitivity gap or a nuisance-recovery guarantee.

4 THEORY AND DIAGNOSTIC EVIDENCE

NCSP should be read as a conditional diagnostic and intervention. The formal assumption does not require every shortcut to be high-frequency. It requires that, for the trained representation and the probe under consideration, nuisance-predictive coordinates occupy a row-wise, head-accessible coordinate support whose perturbation sensitivity is larger than that of the stable semantic alternatives. Frequency can be one mechanism that produces this gap in texture- or background-biased datasets, but it is not required by the method.

The central object is the finite-noise sensitivity in Eq. (2). We therefore treat $S_{c,j}$ as the primary estimand and use gradient quantities only as interpretation. This avoids making the theory depend on global smoothness of the backbone and is better aligned with ReLU networks.

Algorithm 1 Noise-based Class-Conditional Sensitivity Pruning (NCSP)

Input: trained feature extractor ϕ, linear head W, normalized labeled data $\mathcal{D}$, noise scale σ, pruning ratio τ, sensitivity budget N_e.
Freeze ϕ and set $k = \lceil \tau d \rceil$.
for $c = 1$ to C **do**
  Sample $\mathcal{S}_c \subset \{x_i : y_i = c\}$ with $|\mathcal{S}_c| = N_e$.
  Draw an independent $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ for each x and set $T_\sigma(x) = x + \varepsilon$ before the frozen backbone.
  Compute $d_{c,j} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} |\phi_j(T_\sigma(x)) - \phi_j(x)|$ for all $j \in [d]$.
  Set $P_c = \mathrm{TopK}(d_c, k)$ using coordinate index to break ties, and form $M_{c,j} = 0$ for $j \in P_c$, $M_{c,j} = 1$ otherwise.
end for
Refit the head on clean labeled data by solving Eq. (5), enforcing $W \leftarrow W \odot M$ after each optimizer step.
Return: frozen ϕ and masked/refit head $W \odot M$.

Lemma 4.1 (Finite-noise sensitivity for locally Lipschitz features)**.** *Let $\phi_j : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz, let $G \sim \mathcal{N}(0, I_n)$, and define*

$$S_j(x; \sigma) = \mathbb{E}_G \left[|\phi_j(x + \sigma G) - \phi_j(x)| \right].$$

For any x at which $\nabla \phi_j(x)$ exists and the expectations below are finite,

$$S_j(x; \sigma) = \sigma \mathbb{E}_G \left[\left| \int_0^1 \nabla \phi_j(x + t\sigma G)^\top G \, dt \right| \right], \tag{6}$$

where the gradient exists almost everywhere along the segment. If

$$\eta_j(x, \sigma) = \mathbb{E}_G \int_0^1 \|\nabla \phi_j(x + t\sigma G) - \nabla \phi_j(x)\|_2 \, \|G\|_2 \, dt < \infty,$$

then

$$\left| S_j(x; \sigma) - \sqrt{\frac{2}{\pi}} \sigma \, \|\nabla \phi_j(x)\|_2 \right| \le \sigma \eta_j(x, \sigma). \tag{7}$$

For smooth features with a Hessian bound, Eq. (7) recovers the usual $O(n\sigma^2)$ Taylor remainder. For piecewise-affine ReLU networks, the relation is exact inside a fixed activation region, and the only error comes from activation-boundary crossings. Thus the implementation estimates the finite-σ perturbation modulus directly; the local-gradient view is an explanatory corollary, not the definition of the method.

Assumption 4.2 (Axis-aligned head-accessible sensitivity gap)**.** For each class c, let $B_c \subseteq [d]$ denote the row-wise nuisance-linked coordinate support: coordinates through which row c of the linear head can directly use the nuisance factor. The nuisance support is axis-aligned and head-accessible if pruning $W_{c,j}$ for $j \in B_c$ removes the direct nuisance contribution to logit c. We say there is a class-conditional sensitivity gap at budget k if

$$\min_{j \in B_c} S_{c,j} > \max_{j \notin B_c} S_{c,j} + \gamma_c \qquad \text{for some } \gamma_c > 0,$$

with $k \ge |B_c|$. This assumption is used only for nuisance-support recovery. The stability bounds below still hold as near-optimal retained-sensitivity guarantees even when exact nuisance recovery is not available.

Let $\widehat{P}_c = \mathrm{TopK}(d_c, k)$ be the empirical NCSP mask and let $P_c^\star(k) = \mathrm{TopK}(S_c, k)$ be the population top-k sensitivity mask. Define the retained sensitivity mass

$$R_c(P) = \sum_{j \notin P} S_{c,j}.$$

Proposition 4.3 (Finite-sample sensitivity-mask recovery and near-optimality)**.** *Assume that, for all $c \in [C]$ and $j \in [d]$, the centered variables*

$$U_{c,j} - S_{c,j}, \qquad U_{c,j} = |\phi_j(T_\sigma(X)) - \phi_j(X)|, \qquad \textit{conditional on } Y = c,$$

are sub-exponential with parameters (ν, b) uniformly over c, j. Then there is a universal constant $K > 0$ such that, with probability at least $1 - \delta$,

$$\max_{c\in[C],j\in[d]} |d_{c,j} - S_{c,j}| \le \epsilon_N := K \max\left\{ \nu\sqrt{\frac{\log(2Cd/\delta)}{N_e}}, b\frac{\log(2Cd/\delta)}{N_e} \right\}. \tag{8}$$

On this event, the following hold for every class c:

1. *If the population top-k gap*
$$\gamma_c(k) = S_{c,(k)} - S_{c,(k+1)}$$
satisfies $\gamma_c(k) > 2\epsilon_N$, then $\widehat{P}_c = P_c^\star(k)$.

2. *Even without any gap,*
$$R_c(\widehat{P}_c) \le \min_{|P|=k} R_c(P) + 2k\epsilon_N. \tag{9}$$
Thus empirical top-k pruning is near-optimal for removing population sensitivity mass.

3. *Under Assumption 4.2, if $\gamma_c > 2\epsilon_N$ and $k \ge |B_c|$, then $B_c \subseteq \widehat{P}_c$.*

Proposition 4.4 (Convexity and norm control of the constrained refit)**.** *For fixed ϕ and fixed masks $\{\widehat{P}_c\}_{c=1}^C$, Eq.* (5) *is convex in the free entries of W. If $\lambda > 0$, it is strongly convex on the free-coordinate subspace. Moreover, if $\widehat{W}$ is its optimizer, then*

$$\left\|\widehat{W}\right\|_F \le R_\lambda := \sqrt{\frac{2\log C}{\lambda}}. \tag{10}$$

Proposition 4.5 (Post-refit logit-stability bound)**.** *Let $\widehat{W}$ be the masked/refit head returned by Eq.* (5)*, and let*

$$\widehat{z}_c(x) = \langle \widehat{W}_c, \phi(x) \rangle.$$

Then, for every class c,

$$\mathbb{E}[|\widehat{z}_c(T_\sigma(X)) - \widehat{z}_c(X)| \mid Y = c] \le R_\lambda R_c(\widehat{P}_c). \tag{11}$$

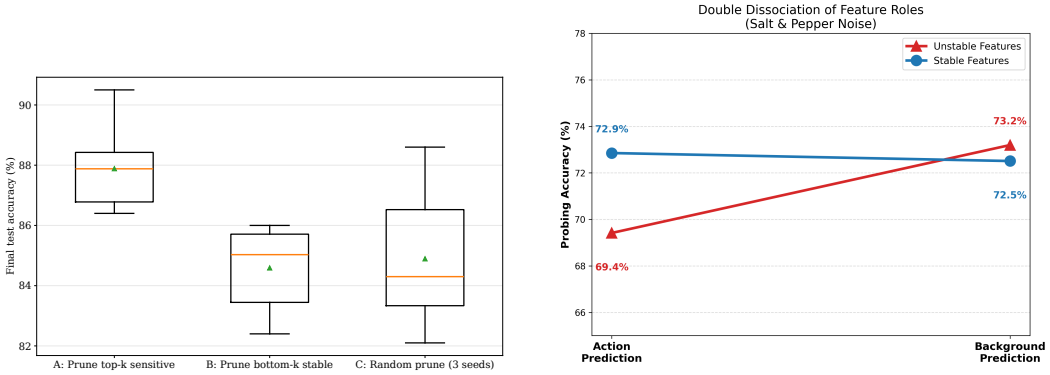
Combining this with Proposition 4.3, with probability at least $1 - \delta$,

$$\mathbb{E}[|\widehat{z}_c(T_\sigma(X)) - \widehat{z}_c(X)| \mid Y = c] \le R_\lambda \left(\min_{|P|=k} R_c(P) + 2k\epsilon_N \right). \tag{12}$$

Thus the NCSP *pipeline, including refitting, has a finite-sample near-optimal stability bound for the retained sensitivity surrogate it optimizes.*

Proposition 4.5 is row-matched: it controls logit row c on examples with $Y = c$. Appendix A.5 gives multiclass prediction and KL statements under cross-row sensitivities and second moments, while Appendix A.6 gives a conditional nuisance-shift invariance result under an additional two-subspace support assumption. These results are conditional. They support group-shift robustness only when the selected row masks cover the head-accessible nuisance support and stable semantic coordinates remain available.

Diagnostic sanity checks. The theory predicts that high-sensitivity coordinates should be useful pruning targets only when they are both removable and nuisance-informative. Figure 1 tests these implications. Reverse pruning checks the retained-sensitivity premise: removing top-sensitive coordinates preserves clean accuracy better than removing bottom-sensitive coordinates, suggesting that the high-sensitivity subset is comparatively removable. Nuisance-probe dissociation checks the nuisance-alignment premise: the sensitive subset carries structured signal more aligned with the confounder than with the label. These checks do not replace group-shift evaluation. They help decide whether the sensitivity-gap and head-accessibility assumptions are plausible on a given model and dataset.



(a) Reverse pruning.

(b) Probe dissociation.

Figure 1: **Sanity checks for the stability–semantics gap.** Left: pruning the most noise-sensitive coordinates preserves clean accuracy better than pruning the least-sensitive coordinates, suggesting that high-sensitivity coordinates are comparatively removable. Right: linear probes trained on sensitive coordinates decode the nuisance/background attribute more strongly than the label, while stable coordinates show the opposite trend.

5 Experiments

Datasets. We evaluate on four spurious-correlation benchmarks. BAR is an action-recognition benchmark where action labels co-occur with typical backgrounds during training, but the test set removes the biased background context (Nam et al., 2020). Waterbirds is a composite benchmark where bird type is spuriously correlated with water or land background (Sagawa et al., 2019). MetaShift constructs natural context shifts from Visual Genome co-occurrences, and we use a cat/dog indoor/outdoor setup with group evaluation (Liang & Zou, 2022; Krishna et al., 2017). FB-CMNIST is a controlled color-bias benchmark with foreground/background color shortcuts and training bias ratio q (Sarridis et al., 2024). Dataset sizes, group distributions, and construction details are in Appendix C.

Protocol. All main experiments use a ResNet-18 backbone initialized with ImageNet weights (He et al., 2016) and a linear classification head. We first train the model with standard ERM, then freeze the backbone and apply NCSP post hoc. Sensitivities are estimated from N_e samples per class; masks are class-conditional; only the linear head is refit. Noise is injected after standard input normalization and before the frozen backbone. We fix $\tau = 0.25$ and $\sigma = 0.2$ across the natural-image benchmarks and use $\sigma = 0.1$ on FB-CMNIST. Checkpoints use label-only validation accuracy; group labels are used only for final reporting. This protocol avoids tuning on test-group metrics but does not guarantee worst-group model selection.

Metrics. For BAR, the standard test split removes the biased background context, so we report average accuracy. For Waterbirds and MetaShift, we report both average accuracy and worst-group accuracy (WGA), where the worst group corresponds to bias-conflicting subpopulations such as waterbirds on land. For FB-CMNIST, we report accuracy on an unbiased balanced test set under increasing training bias ratio q.

5.1 Main results

Table 1 shows that the full NCSP pipeline improves the ERM baseline on the four main benchmarks. On BAR, it reaches 89.64% average accuracy, 29.1 points above ERM and 3.4 points above the strongest listed baseline, EvA-C. The matched controls in Appendix G show, however, that clean-data head refitting explains much of this gain and that sensitivity ranking does not uniformly outperform random or magnitude-based masks.

On Waterbirds, NCSP achieves the best average accuracy in the table (97.11%) and improves WGA over ERM by $+15.7$ points. However, it does not maximize WGA: JTT, DEDIER, and PDE achieve

Table 1: **Main results on real-world spurious-correlation benchmarks.** Mean ± standard deviation over three runs where available. NCSP is a group-label-free post-hoc method. It achieves strong average accuracy and improves WGA over ERM, but it does not dominate methods optimized directly for Waterbirds WGA.

Method	BAR	Waterbirds		MetaShift	
	Avg.	Avg.	Worst	Avg.	Worst
ERM	60.51±4.3	90.10±6.2	63.74±3.2	83.30±0.3	69.60±0.4
LfF (Nam et al., 2020)	62.98±2.8	89.60±2.4	74.98±2.1	80.20±0.3	73.10±1.6
EIIL (Creager et al., 2021)	68.44±1.2	95.88±1.7	77.20±1.0	–	–
JTT (Liu et al., 2021)	68.53±3.2	93.70±0.5	84.98±0.5	90.29±0.5	82.61±0.4
DFR (Kirichenko et al., 2023)	74.12±1.1	91.70±0.7	83.40±1.4	87.86±0.1	79.07±0.1
SiFER (Tiwari & Shenoy, 2023)	72.08±0.4	96.11±0.6	77.22±0.4	–	–
DEDIER (Tiwari et al., 2024)	72.44±0.6	92.20±0.6	89.80±0.5	87.25±1.0	77.82±0.9
PDE (Deng et al., 2023)	72.55±0.4	92.10±0.4	**90.80±0.2**	89.12±0.3	81.25±0.2
ERM+ElRep (Wen et al., 2025)	79.80±0.7	95.50±0.2	60.51±1.6	85.91±0.6	72.14±2.5
EvA-E (He et al., 2025)	73.70±0.8	95.71±1.3	81.31±1.5	81.06±4.9	73.11±5.8
EvA-C (He et al., 2025)	86.22±1.1	91.15±1.3	80.01±1.9	87.28±3.6	81.40±3.4
NCSP (ours)	**89.64±1.02**	**97.11±0.29**	79.43±1.11	**92.49±0.8**	**86.05±1.2**

Table 2: **Unbiased test accuracy on FB-CMNIST under increasing bias.** ERM collapses as $q \to 1$, while NCSP keeps 91.9% accuracy at $q = 0.99$.

Method	$q = 0.90$	$q = 0.95$	$q = 0.99$
ERM	82.5	57.9	25.5
BC-BB (Hong & Yang, 2021)	80.9	66.0	40.9
EnD (Tartaglione et al., 2021)	82.5	57.5	25.7
FLAC (Sarridis et al., 2025)	84.4	63.1	32.4
FairKL (Barbano et al., 2023)	87.6	61.6	42.0
BAdd (Sarridis et al., 2024)	**95.6**	89.0	69.5
NCSP (ours)	94.1	**92.4**	**91.9**

stronger worst-group performance. We therefore interpret Waterbirds as a trade-off result rather than a dominance result. NCSP occupies a high-average-accuracy, low-compute, group-label-free point on the Pareto frontier; methods that explicitly optimize hard examples or worst groups can be preferable when the only objective is Waterbirds WGA.

On MetaShift, NCSP reaches 92.49% average accuracy and 86.05% WGA. On FB-CMNIST, ERM drops from 82.5% to 25.5% as the training bias increases from $q = 0.90$ to $q = 0.99$, whereas NCSP reaches 91.9% at $q = 0.99$. These are results for the complete mask–refit pipeline, not evidence that sensitivity ranking is always the source of the improvement.

5.2 Ablations and diagnostic behavior

Probe choice and operating range. Although the main method uses Gaussian noise, Figure 2 shows related trends for shot and salt-and-pepper probes. This is evidence of limited portability, not probe invariance. We use $\tau = 0.25$ as a fixed operating point rather than a universal optimum. The unaggregated sweep in Appendix G shows stable behavior in a moderate range and clear degradation under aggressive pruning.

Why class conditioning matters. Figure 3a compares the proposed class-conditional mask with global masks. On BAR, class-conditional pruning obtains 89.6% accuracy, whereas global masks reach 86.7%, 85.2%, and 85.5% under white, shot, and salt-and-pepper probes. We expect this gap because BAR classes are associated with different background contexts, so a coordinate can be nuisance-linked for one action but useful for another. Per-row masks allow NCSP to suppress class-specific instabilities while preserving shared stable coordinates.

Prediction stability. Figure 3b measures $D_{\mathrm{KL}}(p(\cdot \mid x) \| p(\cdot \mid T_\sigma(x)))$ during the constrained refit. The pruned/refit head remains below the unpruned baseline level, indicating that the intervention reduces predictive distribution drift under the same probe used for sensitivity estimation. The KL is a diagnostic rather than a training loss, and it follows Proposition 4.5 and Appendix A.5: sensitivity

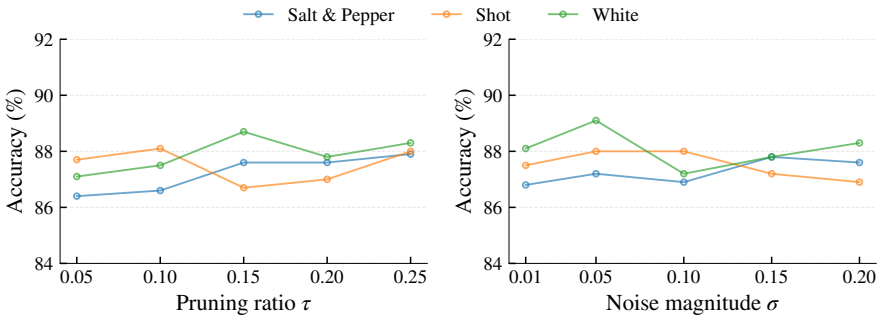
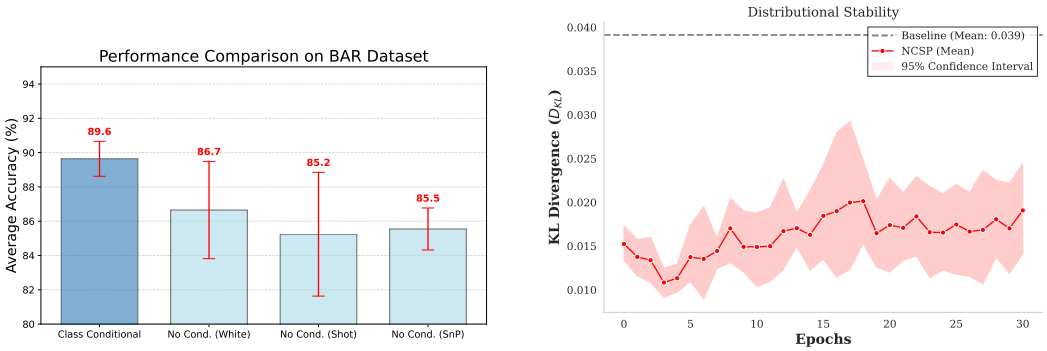


Figure 2: **Probe and hyperparameter behavior.** Accuracy as a function of pruning ratio τ and noise magnitude σ, averaged over the complementary axis. The averaging obscures the high-pruning regime; Appendix G reports the unaggregated pruning sweep.



(a) Class-conditional masks.

(b) Predictive stability.

Figure 3: **Mechanistic ablations.** (a) On BAR, class-conditional masks outperform global masks that remove the same coordinates for all classes. (b) During head refit, NCSP keeps the KL divergence between predictions on clean and perturbed inputs below the unpruned baseline, consistent with the logit-instability bound.

pruning reduces retained coordinate perturbation mass, the ℓ_2-regularized refit controls the norm of the remaining head, and softmax KL is bounded by squared logit drift.

6 Scope, Limitations, and Conclusion

NCSP applies only when nuisance-linked coordinates are probe-sensitive, head-accessible, sufficiently axis-aligned, and separated from stable semantic alternatives. It is also coordinate- and probe-dependent. Orthogonal rotations that preserve the original predictor change the post-intervention result, and a stable nuisance cue can evade the diagnostic entirely.

Reverse pruning, matched random masks, probe agreement, and clean/noisy drift can reject an unsupported intervention; they cannot certify nuisance invariance or worst-group accuracy. Label-only validation can also prefer a high-average, low-WGA model when latent groups are skewed. When group labels or shifted validation data are available, model selection should use them. When they are not, the safe output of the audit may be the original model.

The evidence supports NCSP as a low-cost conditional audit and intervention for frozen visual representations. It does not support sensitivity ranking as a universal substitute for head refitting, group-aware evaluation, or representation retraining.

AI USE STATEMENT

We used AI tools to polish the writing, help draft sections, find spelling mistakes, and fix formatting. All numerical results come from the authors' experiments. The authors checked all text and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies failures caused by spurious visual correlations. Improved average or worst-group accuracy on the reported benchmarks is not a fairness or safety certificate. Deployment in medical, biometric, or other high-stakes settings requires representative subgroup evaluation, confidence intervals, error analysis, monitoring for new failure slices, and direct optimization of the required group objective when suitable annotations are available.

REPRODUCIBILITY STATEMENT

Algorithm 1 specifies the intervention. Appendix B gives tensor shapes, mask enforcement, variants, and complexity; Appendix C gives dataset and evaluation details; Appendix D reports optimization settings; and Appendix A contains the complete proofs. Main results report multi-run statistics where available, while single-run and preliminary stress tests are labeled explicitly.

REFERENCES

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

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International conference on machine learning*, pp. 528–539. PMLR, 2020.

Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

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

Preetam Prabhu Srikar Dammu and Chirag Shah. Detecting spurious correlations via robust visual concepts in real and ai-generated image classification. *arXiv preprint arXiv:2311.01655*, 2023.

Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems*, 36:1390–1402, 2023.

Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1962–1981, 2022.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

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

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Qiyuan He, Kai Xu, and Angela Yao. Eva: Erasing spurious correlations with activations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=zKvrOOBouT`.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising Bias in Compressed Models. *arXiv preprint arXiv:2010.03058*, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

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.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

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

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pp. 598–605, 1989.

Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MTex8qKavoS.

Evan Zheran 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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 6781–6792, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Vito Paolo Pastore, Massimiliano Ciranni, Davide Marinelli, Francesca Odone, and Vittorio Murino. Looking at model debiasing through the lens of anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2548–2557, 2025.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with XRM. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40551–40569. PMLR, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Badd: Bias mitigation through bias addition. *arXiv preprint arXiv:2408.11439*, 2024.

Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. FLAC: Fairness-Aware Representation Learning by Suppressing Attribute-Class Associations . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(02):1148–1160, February 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3487254. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2024.3487254.

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

Nimit S. Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352, 2020.

Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13508–13517, 2021.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34330–34343. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/tiwari23a.html.

Rishabh Tiwari, Durga Sivasubramanian, Anmol Mekala, Ganesh Ramakrishnan, and Pradeep Shenoy. Using early readouts to mediate featural bias in distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2638–2647, January 2024.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8684–8694, 2020.

Tao Wen, Zihan Wang, Quan Zhang, and Qi Lei. Elastic representation: Mitigating spurious correlations for group robustness. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 541–549. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/wen25a.html.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Samira Zare and Hien Van Nguyen. Frustratingly easy environment discovery for invariant learning. *Computer Sciences & Mathematics Forum*, 9(1):2, 2024. doi: 10.3390/cmsf2024009002.

Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang, and Zicong Hong. Amend to alignment: decoupled prompt tuning for mitigating spurious correlation in vision-language models. In *Forty-first International Conference on Machine Learning*, 2024.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. NICO++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16036–16047, 2023.

A FULL THEORY AND PROOFS

This appendix gives the formal details behind the theory statements in Section 4. The results give a conditional account of NCSP. The method always estimates class-conditional finite-noise sensitivity. When nuisance-linked coordinates are head-accessible, approximately axis-aligned, and more sensitive to the chosen perturbation probe than stable semantic alternatives, the resulting masks can remove direct shortcut reliance from the final head. The theory does not claim that all spurious features are high-frequency, nor that all semantic features are stable.

A.1 FINITE-NOISE SENSITIVITY FOR NONSMOOTH NETWORKS

Proof of Lemma 4.1. Let $G \sim \mathcal{N}(0, I_n)$ and fix x. Because ϕ_j is locally Lipschitz, Rademacher's theorem implies that it is differentiable almost everywhere. For almost every Gaussian direction G, the one-dimensional function

$$h_G(t) = \phi_j(x + t\sigma G), \qquad t \in [0, 1],$$

is absolutely continuous, and therefore

$$\phi_j(x + \sigma G) - \phi_j(x) = h_G(1) - h_G(0) = \sigma \int_0^1 \nabla\phi_j(x + t\sigma G)^\top G \, dt.$$

Taking absolute values and expectation gives Eq. (6). At a point x where $\nabla\phi_j(x)$ exists, write

$$\sigma \int_0^1 \nabla\phi_j(x + t\sigma G)^\top G \, dt = \sigma \nabla\phi_j(x)^\top G + \sigma \int_0^1 \big(\nabla\phi_j(x + t\sigma G) - \nabla\phi_j(x)\big)^\top G \, dt.$$

Using $\big||a + b| - |a|\big| \le |b|$ and Cauchy–Schwarz,

$$\begin{aligned} &\big|S_j(x;\sigma) - \mathbb{E}_G \left|\sigma \nabla\phi_j(x)^\top G\right|\big| \\ &\qquad \le \sigma \mathbb{E}_G \int_0^1 \left\|\nabla\phi_j(x + t\sigma G) - \nabla\phi_j(x)\right\|_2 \|G\|_2 \, dt = \sigma\eta_j(x,\sigma). \end{aligned}$$

Since $\nabla\phi_j(x)^\top G$ is a centered Gaussian with variance $\|\nabla\phi_j(x)\|_2^2$,

$$\mathbb{E}_G \left|\sigma \nabla\phi_j(x)^\top G\right| = \sqrt{\frac{2}{\pi}}\sigma \left\|\nabla\phi_j(x)\right\|_2.$$

Combining the last two displays proves Eq. (7). □

Smooth special case. If ϕ_j is twice differentiable and $\left\|\nabla^2\phi_j(u)\right\|_{\mathrm{op}} \le H_j$ on the region reached by the perturbation, then

$$\left\|\nabla\phi_j(x + t\sigma G) - \nabla\phi_j(x)\right\|_2 \le H_j t\sigma \|G\|_2.$$

Therefore

$$\eta_j(x,\sigma) \le H_j \sigma \mathbb{E}\|G\|_2^2 \int_0^1 t \, dt = \frac{H_j}{2}\sigma n,$$

and Lemma 4.1 gives

$$\left|S_j(x;\sigma) - \sqrt{\frac{2}{\pi}}\sigma \left\|\nabla\phi_j(x)\right\|_2\right| \le \frac{H_j}{2} n\sigma^2.$$

Thus the previous Taylor-style bound is recovered as a special case.

Piecewise-affine ReLU special case. Suppose ϕ_j is L-Lipschitz and piecewise affine, and suppose x is not on an activation boundary. Let a_x be the affine gradient in the activation region containing x, and let E_{cross} be the event that the segment $\{x + t\sigma G : t \in [0, 1]\}$ crosses an activation boundary. On E_{cross}^c,

$$\phi_j(x + \sigma G) - \phi_j(x) = \sigma a_x^\top G.$$

Hence

$$\left| S_j(x;\sigma) - \sqrt{\frac{2}{\pi}}\sigma \left\| a_x \right\|_2 \right|$$
$$\leq \mathbb{E}\left[\left| |\phi_j(x+\sigma G) - \phi_j(x)| - \left|\sigma a_x^\top G\right| \right| \mathbf{1}\{E_{\text{cross}}\} \right]$$
$$\leq 2L\sigma \mathbb{E}\left[\|G\|_2 \mathbf{1}\{E_{\text{cross}}\} \right] \leq 2L\sigma \sqrt{n\, \mathbb{P}(E_{\text{cross}})}.$$

Thus for ReLU networks the local-gradient relation is exact within a fixed activation region, with an error controlled by the probability of crossing a boundary.

Interpretation. The estimand is the expected absolute scalar deviation

$$S_j(x;\sigma) = \mathbb{E}_\varepsilon \left| \phi_j(x+\varepsilon) - \phi_j(x) \right|.$$

The implementation estimates this finite-noise quantity directly using elementwise absolute differences. The gradient formula explains the behavior of the quantity under additional regularity; it is not required for defining NCSP.

A.2 Finite-sample concentration and retained-sensitivity optimality

Proof of Proposition 4.3. Fix a class c and coordinate j. Let

$$U_i^{(c,j)} = |\phi_j(T_\sigma(X_i)) - \phi_j(X_i)|$$

for i.i.d. samples conditional on $Y = c$. Then

$$d_{c,j} = \frac{1}{N_e} \sum_{i=1}^{N_e} U_i^{(c,j)}.$$

Bernstein's inequality for sub-exponential variables and a union bound over the Cd class–coordinate pairs give

$$\max_{c\in[C], j\in[d]} |d_{c,j} - S_{c,j}| \leq K \max\left\{ \nu\sqrt{\frac{\log(2Cd/\delta)}{N_e}}, b\frac{\log(2Cd/\delta)}{N_e} \right\}$$

with probability at least $1-\delta$, which is Eq. (8).

Now work on the event

$$\max_{c,j} |d_{c,j} - S_{c,j}| \leq \epsilon_N.$$

For exact top-k recovery, let $S_{c,(1)} \geq \cdots \geq S_{c,(d)}$ denote the sorted sensitivities. If

$$S_{c,(k)} - S_{c,(k+1)} > 2\epsilon_N,$$

then every population top-k coordinate has empirical sensitivity larger than every non-top-k coordinate, so $\widehat{P}_c = P_c^\star(k)$.

For near-optimality, let

$$P_c^\star(k) \in \arg\min_{|P|=k} R_c(P)$$

be a population top-k set. Equivalently, $P_c^\star(k)$ maximizes $\sum_{j\in P} S_{c,j}$ among all sets of size k. Since $\widehat{P}_c$ maximizes empirical sensitivity mass,

$$\sum_{j\in\widehat{P}_c} d_{c,j} \geq \sum_{j\in P_c^\star(k)} d_{c,j}.$$

Using uniform concentration on both sets,

$$\sum_{j\in\widehat{P}_c} S_{c,j} \geq \sum_{j\in\widehat{P}_c} d_{c,j} - k\epsilon_N \geq \sum_{j\in P_c^\star(k)} d_{c,j} - k\epsilon_N \geq \sum_{j\in P_c^\star(k)} S_{c,j} - 2k\epsilon_N.$$

Subtracting both sides from $\sum_{j=1}^d S_{c,j}$ gives

$$R_c(\widehat{P}_c) \leq R_c(P_c^\star(k)) + 2k\epsilon_N = \min_{|P|=k} R_c(P) + 2k\epsilon_N.$$

Finally, if Assumption 4.2 holds with $\gamma_c > 2\epsilon_N$, then for every $j \in B_c$ and $\ell \notin B_c$,

$$d_{c,j} \geq S_{c,j} - \epsilon_N > S_{c,\ell} + \epsilon_N \geq d_{c,\ell}.$$

Thus every coordinate in B_c ranks above every coordinate outside B_c. Since $k \geq |B_c|$, all coordinates in B_c are included in $\widehat{P}_c$. □

Bounded special case. If $U_{c,j} \in [0, B]$ almost surely, Hoeffding's inequality gives the sharper radius

$$\epsilon_N = B\sqrt{\frac{\log(2Cd/\delta)}{2N_e}}.$$

The bounded result is a special case and is not inferred from Gaussian noise, which has unbounded support. A global Lipschitz condition on ϕ_j instead controls $U_{c,j}$ by $L_j \|\varepsilon\|_2$ and supplies the light-tail condition used in the main proposition.

What the proposition does and does not imply. The proposition first guarantees recovery of the population sensitivity mask, not recovery of nuisance coordinates by itself. Nuisance recovery additionally requires an alignment condition such as Assumption 4.2. Without such an alignment condition, no sensitivity-only ranking method can guarantee nuisance recovery; see Appendix A.7.

A.3 Convexity and norm control of the constrained refit

Proof of Proposition 4.4. For fixed features $h_i = \phi(x_i)$, the multiclass logits Wh_i are affine in W. The softmax cross-entropy is convex in its logits, and composition with an affine map preserves convexity. The feasible set

$$\{W : W_{c,j} = 0 \ \ \forall c \in [C], \ j \in \widehat{P}_c\}$$

is an affine subspace. Hence the constrained refit objective is convex on the feasible set. Adding $\lambda \|W\|_F^2 / 2$ makes the objective strongly convex in the free coordinates whenever $\lambda > 0$.

For the norm bound, note that $W = 0$ is feasible. At $W = 0$, the softmax distribution is uniform, so the average cross-entropy is $\log C$. Since $\widehat{W}$ minimizes the constrained objective,

$$\frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{ce}}(\widehat{W}\phi(x_i), y_i) + \frac{\lambda}{2}\left\|\widehat{W}\right\|_F^2 \leq \log C.$$

The cross-entropy term is nonnegative, so

$$\frac{\lambda}{2}\left\|\widehat{W}\right\|_F^2 \leq \log C.$$

Rearranging gives

$$\left\|\widehat{W}\right\|_F \leq \sqrt{\frac{2\log C}{\lambda}} = R_\lambda.$$

□

A.4 Post-refit logit stability

Proof of Proposition 4.5. Let

$$\Delta\phi(x) = \phi(T_\sigma(x)) - \phi(x).$$

Because $\widehat{W}_{c,j} = 0$ for all $j \in \widehat{P}_c$,

$$\begin{aligned}
|\widehat{z}_c(T_\sigma(x)) - \widehat{z}_c(x)| &= \left|\widehat{W}_c^\top \Delta\phi(x)\right| \\
&= \left|\sum_{j \notin \widehat{P}_c} \widehat{W}_{c,j}\Delta\phi_j(x)\right| \\
&\leq \left\|\widehat{W}_c\right\|_\infty \sum_{j \notin \widehat{P}_c} |\Delta\phi_j(x)| \\
&\leq R_\lambda \sum_{j \notin \widehat{P}_c} |\Delta\phi_j(x)|,
\end{aligned}$$

where the last inequality uses $\left\|\widehat{W}_c\right\|_\infty \le \left\|\widehat{W}\right\|_F \le R_\lambda$ from Proposition 4.4. Taking expectation conditional on $Y = c$ gives

$$\begin{aligned}\mathbb{E}[|\widehat{z}_c(T_\sigma(X)) - \widehat{z}_c(X)| \mid Y = c] &\le R_\lambda \sum_{j \notin \widehat{P}_c} \mathbb{E}[|\Delta\phi_j(X)| \mid Y = c] \\ &= R_\lambda \sum_{j \notin \widehat{P}_c} S_{c,j} = R_\lambda R_c(\widehat{P}_c).\end{aligned}$$

The finite-sample near-optimal bound follows by substituting Eq. (9). □

Relation to fixed-head pruning. For a fixed, non-refit head, the triangle inequality gives

$$|z_c(T_\sigma(x)) - z_c(x)| \le \sum_{j=1}^{d} |W_{c,j}|\,|\Delta\phi_j(x)|\,,$$

and pruning row-c coordinates P_c removes the corresponding terms. This fixed-head bound is useful intuition, but it would suggest ranking by $|W_{c,j}|\,S_{c,j}$. Proposition 4.5 is better aligned with NCSP: after the ℓ_2-regularized refit, the head norm is controlled, so pruning by sensitivity alone yields a post-refit retained-sensitivity guarantee.

A.5 Prediction and KL stability

For an input class y and representation coordinate j, define

$$S_{y,j} = \mathbb{E}[|\Delta\phi_j(X)| \mid Y = y], \qquad V_{y,j} = \mathbb{E}[(\Delta\phi_j(X))^2 \mid Y = y].$$

Proposition A.1 (Cross-row expected stability)**.** *For input class y and logit row r, the masked/refit head satisfies*

$$\mathbb{E}[|\Delta z_r(X)| \mid Y = y] \le R_\lambda \sum_{j \notin \widehat{P}_r} S_{y,j}. \tag{13}$$

Moreover,

$$\mathbb{E}[D_{\mathrm{KL}}(p(\cdot \mid X) \| p(\cdot \mid T_\sigma(X))) \mid Y = y] \le \frac{R_\lambda^2}{2} \sum_{r=1}^{C} \sum_{j \notin \widehat{P}_r} V_{y,j}. \tag{14}$$

Thus multiclass stability on class-y inputs requires control of every row mask under the class-y input distribution; it does not follow from the row-matched bound alone.

Corollary A.2 (Pointwise prediction and KL stability)**.** *For a masked/refit head W, define the class-c logit drift*

$$\delta_c(x) = |z_c(T_\sigma(x)) - z_c(x)|\,.$$

If y is the clean predicted class and

$$z_y(x) - \max_{r \ne y} z_r(x) > \delta_y(x) + \max_{r \ne y} \delta_r(x),$$

then the prediction is unchanged by the perturbation T_σ. Moreover, for

$$p(\cdot \mid x) = \mathrm{softmax}(z(x)), \qquad q(\cdot \mid x) = \mathrm{softmax}(z(T_\sigma(x))),$$

$$D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, q(\cdot \mid x)) \le \frac{1}{2}\,\|z(T_\sigma(x)) - z(x)\|_2^2\,. \tag{15}$$

Finally, the sensitivity theory can be connected to group robustness only under an additional representation assumption.

Proof of Corollary A.2. Let $y = \arg\max_c z_c(x)$ be the clean predicted class. For any competitor $r \ne y$,

$$z_y(T_\sigma(x)) \ge z_y(x) - \delta_y(x), \qquad z_r(T_\sigma(x)) \le z_r(x) + \delta_r(x).$$

Therefore $z_y(T_\sigma(x)) > z_r(T_\sigma(x))$ for all $r \ne y$ whenever

$$z_y(x) - z_r(x) > \delta_y(x) + \delta_r(x) \qquad \forall r \ne y.$$

The stated sufficient condition implies this inequality for every competitor, so the prediction is unchanged.

For the KL bound, let

$$F(z) = \log \sum_{c=1}^{C} e^{z_c}.$$

The softmax distribution is $\nabla F(z)$. If

$$p = \mathrm{softmax}(z), \qquad q = \mathrm{softmax}(z'),$$

then

$$D_{\mathrm{KL}}(p\|q) = F(z') - F(z) - \nabla F(z)^\top (z' - z).$$

This is the Bregman divergence of F. The Hessian of F is the covariance matrix of a categorical distribution, whose operator norm is at most 1. Hence F is 1-smooth, and

$$F(z') - F(z) - \nabla F(z)^\top (z' - z) \le \frac{1}{2} \|z' - z\|_2^2 .$$

Taking $z' = z(T_\sigma(x))$ proves Eq. (15). □

Proof of Proposition A.1. Equation (13) follows from the same triangle inequality as Proposition 4.5, conditioning on $Y = y$ while retaining row r. For the KL bound, apply Eq. (15) and, for every row r,

$$(\Delta z_r)^2 = (\widehat{W}_r^\top \Delta\phi)^2 \le \left\|\widehat{W}_r\right\|_2^2 \sum_{j \notin \widehat{P}_r} (\Delta\phi_j)^2 \le R_\lambda^2 \sum_{j \notin \widehat{P}_r} (\Delta\phi_j)^2.$$

Summing over rows and taking expectation conditional on $Y = y$ proves Eq. (14). □

A.6 Conditional invariance under nuisance shifts

Proposition A.3 (Conditional invariance under nuisance shifts)**.** *Let $X = g(S, Z)$, where $Y = h(S)$ and Z is non-causal for the target. Suppose the frozen representation decomposes as*

$$\phi(g(S, Z)) = \phi^{\mathrm{sem}}(S) + \phi^{\mathrm{bias}}(Z) + \xi(S, Z),$$

and suppose that for each row c, the nuisance component used by that row is supported on B_c. If $B_c \subseteq \widehat{P}_c$ for all c and $\xi = 0$, then every feasible masked head has logits that depend only on S. Consequently, for any two distributions P and Q satisfying $P(S, Y) = Q(S, Y)$ but allowing arbitrary shifts in $P(Z \mid S, Y)$, the population loss of any feasible masked head is identical under P and Q. If instead $\max_c \left|W_c^\top \xi(S, Z)\right| \le \eta$, then the masked classifier can differ from the ideal semantic classifier only on examples whose semantic margin is at most 2η.

Proof of Proposition A.3. For a feasible masked head, $W_{c,j} = 0$ for every $j \in \widehat{P}_c$. If $B_c \subseteq \widehat{P}_c$, then row c cannot use any coordinate in its nuisance support. When $\xi = 0$,

$$\phi(g(S, Z)) = \phi^{\mathrm{sem}}(S) + \phi^{\mathrm{bias}}(Z).$$

By the support assumption, the direct row-c nuisance contribution is supported only on B_c, and those entries are zeroed by the mask. Thus

$$W_c^\top \phi(g(S, Z)) = W_c^\top \phi^{\mathrm{sem}}(S)$$

for every class c. All logits therefore depend only on S, not on Z.

Let P and Q be two distributions with the same marginal distribution over (S, Y) but different nuisance conditionals. Since the logits and loss depend only on (S, Y), the population loss under P equals the population loss under Q.

For the approximate case, write

$$z_c(S, Z) = z_c^{\mathrm{sem}}(S) + e_c(S, Z), \qquad |e_c(S, Z)| \le \eta.$$

If the ideal semantic classifier is correct with margin

$$m_{\mathrm{sem}}(S, Y) = z_Y^{\mathrm{sem}}(S) - \max_{r \neq Y} z_r^{\mathrm{sem}}(S) > 2\eta,$$

then

$$z_Y(S, Z) \geq z_Y^{\mathrm{sem}}(S) - \eta > \max_{r \neq Y} z_r^{\mathrm{sem}}(S) + \eta \geq \max_{r \neq Y} z_r(S, Z),$$

so the masked classifier is also correct. Hence the masked classifier can differ from the ideal semantic classifier only on examples with semantic margin at most 2η. □

Interpretation. Proposition A.3 is a conditional result. It says that if sensitivity pruning covers the row-wise head-accessible nuisance support and stable semantic coordinates remain available, then the final head no longer directly depends on the nuisance factor. This is the formal version of the two-subspace story: sensitivity identifies a removable nuisance support; constrained refitting recovers a classifier in the remaining semantic subspace.

A.7 A NECESSARY IDENTIFIABILITY CONDITION

The sensitivity-gap assumption is needed beyond the proof. Some alignment between sensitivity and nuisance structure is necessary for any method that observes only coordinate sensitivities.

Proposition A.4 (No sensitivity-only nuisance recovery without alignment)**.** *Consider two coordinates and a pruning budget $k = 1$. Suppose a pruning rule observes only the sensitivity vector $(S_{c,1}, S_{c,2})$. If $S_{c,1} = S_{c,2}$, and there are two possible representations consistent with this same sensitivity vector—one in which coordinate 1 is the nuisance coordinate and one in which coordinate 2 is the nuisance coordinate—then no deterministic sensitivity-only rule can recover the nuisance coordinate in both representations. Any randomized rule fails with probability at least $1/2$ in one of the two representations.*

Proof. Because the two representations have the same observed sensitivity vector, the pruning rule must produce the same output distribution in both cases. A deterministic rule chooses either coordinate 1 or coordinate 2; whichever coordinate it chooses, it fails in the representation where the other coordinate is the nuisance coordinate. For a randomized rule, let p be the probability of choosing coordinate 1. The failure probability is $1 - p$ in the representation where coordinate 1 is nuisance and p in the representation where coordinate 2 is nuisance. The maximum of these two failure probabilities is at least $1/2$. □

This lower bound explains the scope of NCSP: sensitivity can identify nuisance coordinates only when the learned representation creates a measurable sensitivity–nuisance alignment, such as the gap tested by the diagnostics in Appendix E.

A.8 FREQUENCY IS ONE POSSIBLE MECHANISM, NOT THE ASSUMPTION

Lemma 4.1 links finite-noise sensitivity to path-integrated input-gradient response. In background- or texture-biased vision datasets, nuisance cues are often spatially localized, texture-like, or high-frequency, which can produce larger gradients than smoother object-level evidence (Geirhos et al., 2018; Ilyas et al., 2019; Yin et al., 2019; Wang et al., 2020). A sinusoidal example illustrates this mechanism: if $f(t) = \sin(kt)$, then $|f'(t)| = |k \cos(kt)|$, so gradient magnitude scales with frequency.

This explains why a sensitivity gap often appears in the benchmarks studied here. NCSP does not hard-code it as a prior. The method ranks coordinates by measured class-conditional brittleness in the trained representation. It can therefore succeed when the nuisance is a brittle foreground feature, and it can fail when the nuisance is stable or when semantic evidence is itself sensitive. Appendix F.4 provides a role-reversal stress test supporting this interpretation.

B ALGORITHMIC DETAILS, VARIANTS, AND COMPLEXITY

B.1 TENSOR SHAPES AND SENSITIVITY ESTIMATION

For a batch $X_b \in \mathbb{R}^{B\times C_{\text{in}}\times H\times W}$, NCSP computes clean features

$$H_b = \phi(X_b) \in \mathbb{R}^{B\times d}$$

and perturbed features

$$H_b^\sigma = \phi(T_\sigma(X_b)) \in \mathbb{R}^{B\times d}.$$

It accumulates

$$\sum_{x\in S_c} |\phi(T_\sigma(x)) - \phi(x)|$$

separately for each class c and divides by the number of sampled examples in that class. For ResNet-18, $d = 512$ after global average pooling. For ResNet-50, $d = 2048$. For ViT-B/16, we apply the same coordinate-level procedure to the pooled representation used by the linear classification head.

Noise is injected after standard input normalization and before the frozen backbone. Unless otherwise stated, the main experiments use one noisy copy per clean example. Multiple noisy copies can reduce estimator variance, but we did not find this necessary for the reported budgets.

B.2 MASK ENFORCEMENT

During head refit, masked entries are kept exactly zero. In implementation, after each optimizer step we multiply the head by the binary free-coordinate mask,

$$W \leftarrow W \odot M,$$

where $M_{c,j} = 0$ for $j \in P_c$ and $M_{c,j} = 1$ otherwise. This projected update is equivalent to optimizing over the constrained parameterization in which masked entries are removed.

B.3 FIXED AND ADAPTIVE PRUNING BUDGETS

The main experiments use a fixed pruning ratio $\tau = 0.25$ for simplicity and comparability across datasets. The fixed-ratio version is useful as a transparent baseline: every class prunes exactly $k = \lceil \tau d \rceil$ head connections.

We also evaluate an adaptive variant that sets the pruning ratio from the dispersion of the class-conditional sensitivity distribution. Let

$$\text{CV}_c = \frac{\text{std}(d_c)}{\text{mean}(d_c) + \epsilon},$$

where $\epsilon > 0$ is a small numerical constant. The adaptive ratio is

$$\tau_c = \text{clip}\big(\alpha\, \text{CV}_c, \tau_{\min}, \tau_{\max}\big), \qquad k_c = \lceil \tau_c d \rceil.$$

This rule increases pruning only when the class-c sensitivity distribution has a pronounced unstable tail. When the distribution is tight, τ_c remains small, indicating that the probe has not exposed a clear brittle subspace. We use the fixed $\tau = 0.25$ setting in the main paper and report the adaptive variant as a robustness check in Appendix F.5.

A safety-oriented extension may instead set $k_c = 0$ and return the original row when the estimated top-k gap does not exceed its concentration radius or when top-sensitive removal is no safer on label-only validation data than matched random and bottom-sensitive controls. This abstaining rule is a recommended screening procedure, not the fixed-ratio method evaluated in the main tables.

B.4 COMPUTATIONAL COMPLEXITY

Sensitivity estimation uses two frozen-backbone forward passes per sampled example: one clean and one perturbed. For C classes and budget N_e per class, the additional cost is $2CN_e$ frozen-backbone

Table B.1: Annotation and training footprint of representative debiasing approaches. "Group labels" refers to use of nuisance/group annotations for training or model selection, not final evaluation.

Method	Group labels for debiasing?	Retrains backbone?	Aux. model or env. discovery?	Primary intervention
Group DRO (Sagawa et al., 2019)	Yes	Yes	No	Worst-group objective
JTT (Liu et al., 2021)	No	Yes	Uses error set	Sample reweighting/retraining
DFR (Kirichenko et al., 2023)	Uses group-balanced val	No	No	Last-layer retraining
FEED (Zare & Nguyen, 2024)	No	Yes	Environment discovery	Invariant learning pipeline
XRM (Pezeshki et al., 2024)	No	Modifies training/refit pipeline	Environment discovery	Cross-risk minimization
Anomaly-based debiasing (Pastore et al., 2025)	No	Method-dependent	Anomaly/slice discovery	Sample-level bias discovery
NCSP	No	No	No	Coordinate-level head masking and refit

forward passes. The refit optimizes only the linear head. With pruning ratio τ, the number of free head parameters is

$$C(d - \lceil \tau d \rceil).$$

Memory usage is dominated by standard forward activations and the $C \times d$ sensitivity and mask matrices. All reported experiments used one NVIDIA RTX A4000 GPU. The post-hoc stage requires paired frozen-backbone forward passes and optimization of the linear head; it does not backpropagate through the backbone.

B.5 Annotation and training-footprint comparison

Table B.1 summarizes the design-space distinction between NCSP and representative label-free or weakly supervised debiasing methods. The table complements the empirical comparison. It shows that NCSP is a post-hoc coordinate-level intervention, whereas methods such as JTT, FEED, XRM, and anomaly-detection approaches typically act at the sample, environment, or training-pipeline level.

C Dataset statistics and evaluation details

Waterbirds Waterbirds combines CUB bird photographs with Places backgrounds. The label is bird type and the nuisance attribute is background. The training distribution is highly skewed, while validation and test sets support group evaluation. Group labels are used only for reporting average and worst-group accuracy.

Table C.1: Waterbirds group statistics. Minority groups are bias-conflicting groups.

Label Y	Background Z	Train	Val	Test
Landbird	Land	3498	467	2255
Landbird	Water	184	466	2255
Waterbird	Land	56	133	642
Waterbird	Water	1057	133	642
Total	–	4795	1199	5794

BAR BAR contains six action classes, each correlated with a typical background context during training. The test set removes the biased backgrounds, so average test accuracy is the primary metric and worst-group accuracy is not well-defined under the standard split.

Table C.2: BAR action classes and associated biased contexts.

Action class	Bias-aligned context	Total images
Climbing	Rock wall	326
Diving	Underwater	520
Fishing	Water surface	163
Racing	Racetrack	336
Throwing	Grass field	317
Vaulting	Gym indoor	279

MetaShift We use the cat-versus-dog subset with indoor/outdoor context shift. The test set is balanced across subpopulations to support worst-group evaluation. Group labels are used only for reporting.

FB-CMNIST FB-CMNIST introduces foreground and background color shortcuts with controllable training bias ratios $q \in \{0.90, 0.95, 0.99\}$ and evaluates on an unbiased test set.

C.1 Dataset sizes

Table C.3: Dataset sizes used in the experiments.

Dataset	Classes	Train	Val	Test
Waterbirds	2	4795	1199	5794
BAR	6	1941	327	327
MetaShift	2	1608	288	288
FB-CMNIST	10	60000	12000	10000

C.2 CelebA scope-boundary experiment

CelebA is included as a scope-boundary experiment rather than a main benchmark. The target is blond hair and the nuisance attribute is gender. This setting is challenging for NCSP because the gender–hair-color shortcut may be encoded deeply in the backbone and may not appear as a cleanly separable final-layer brittle subspace. We therefore use CelebA to clarify failure modes, not to claim universal subpopulation-shift robustness. Results are reported in Appendix F.6.

D Hyperparameters, optimization, and reproducibility

Training and refit protocol. All main experiments begin with standard ERM training. During this initial ERM phase, the backbone is trained end-to-end. NCSP is applied post hoc: the backbone is frozen, sensitivities are estimated from paired clean/noisy forward passes, row-specific masks are formed, and only the linear classification head is re-optimized under the fixed mask constraints.

Model selection. Checkpoints are selected using label-only validation accuracy. For datasets with group annotations, group labels are used only for final reporting of group metrics, not for sensitivity estimation, pruning, head refit, or model selection. Thus "group-label-free" means free of nuisance/group labels during the debiasing procedure; it does not mean label-free classification. As Table G.2 shows, this rule can miss the best WGA when latent groups are skewed. Fixed hyperparameters avoid test-group tuning but do not protect hidden groups. Group-aware or shifted validation should be used whenever it is available.

E Diagnostic protocols and additional analyses

E.1 Reverse-pruning and nuisance-probe diagnostics

The sanity checks in Figure 1 test whether the sensitivity-gap condition is plausible for a trained representation. They are diagnostic analyses, not part of the NCSP training pipeline.

Reverse pruning. Coordinates are ranked by class-conditional sensitivity. We compare three interventions with the same pruning budget: pruning the top-k most sensitive coordinates, pruning the bottom-k least sensitive coordinates, and pruning k random coordinates. If top-sensitive pruning preserves clean accuracy better than bottom-sensitive pruning, then the high-sensitivity subset is comparatively removable.

Probe dissociation. We train simple linear probes on high-sensitivity and low-sensitivity coordinate subsets to predict the target label Y and nuisance attribute Z. The intended diagnostic pattern is a double dissociation: sensitive coordinates predict Z better than Y, while stable coordinates predict Y better than Z. This uses nuisance labels only for post-hoc analysis of the learned representation; NCSP itself does not use Z.

Table D.1: Hyperparameters used for NCSP. The backbone is frozen during sensitivity estimation and constrained head refit.

Parameter	Waterbirds	BAR	MetaShift	FB-CMNIST
Backbone	ResNet-18	ResNet-18	ResNet-18	ResNet-18
Optimizer for head refit	Adam	SGD	Adam	SGD
Learning rate	0.01	0.01	0.001	0.01
Weight decay	10^{-4}	10^{-4}	10^{-4}	5×10^{-4}
Momentum	–	0.9	0.9	0.9
Batch size	64	128	128	128
Refit epochs	30	30	50	30
Sensitivity budget N_e	100	100	500	100
Pruning ratio τ	0.25	0.25	0.25	0.25
Noise level σ	0.2	0.2	0.2	0.1

Table E.1: Coordinate-level sensitivity diagnostics. Purity is measured by independent single-coordinate nuisance probes using a chance $+5\%$ threshold. The dissociation gap summarizes the first coordinate in the sensitivity ordering: positive values indicate nuisance preference for sensitive coordinates and label preference for stable coordinates.

Dataset / backbone	d	k	τ	Sensitive-coordinate nuisance purity	Gini of sensitivity	Dissociation gap at $j = 1$
BAR / ResNet-18	512	128	0.25	100% (128/128)	0.117	+41%
Waterbirds / ResNet-50	2048	512	0.25	100% (512/512)	0.121	+38%

E.2 Coordinate purity and signal concentration

The sensitivity signal might be driven by only a few outlier coordinates, with many selected dimensions being false positives. We therefore analyze selected coordinates directly.

For each selected coordinate, we train a single-feature logistic probe to predict the nuisance attribute Z and count the coordinate as nuisance-predictive if probe accuracy exceeds chance by at least 5%. We also report the Gini coefficient of the full sensitivity distribution. Low Gini with high selected-coordinate purity indicates a broad brittle subspace rather than a single spiking coordinate.

Purity is stable across prefix sizes. On BAR/ResNet-18, the purity trend is stable as the prefix size varies with the pruning budget. At $k = 64$, purity remains 100%; at $k = 128$, purity remains 100%; and at $k = 192$, purity is 98.4%. Thus the result is not an artifact of selecting exactly 128 coordinates. The opposite Y/Z dissociation between sensitive and stable subsets adds evidence beyond the high nuisance predictiveness of the selected coordinates.

E.3 Attribution and sensitivity

We compare global sensitivity with feature attribution scores. Sensitivity measures something different from attribution, even when coordinates cannot be semantically named: some high-attribution coordinates are stable, while others are unstable. NCSP targets the high-sensitivity subset and avoids relying on unsupported semantic labels for individual coordinates.

Note on qualitative visualizations. We do not rely on claims that individual feature dimensions have human-interpretable names. If Grad-CAM or feature-map visualizations are included, they should be presented as illustrative evidence only. The core evidence for NCSP is quantitative: reverse pruning, nuisance-probe dissociation, coordinate purity, predictive-stability reduction, and downstream group-shift performance.

E.4 Spectral analysis

Spectral analysis provides one empirical mechanism for the sensitivity gap in background- and texture-biased datasets. We rank final-layer channels by sensitivity, compute 2D Fourier spectra of pre-pooling activation maps, and compare radial power spectral density for stable and unstable

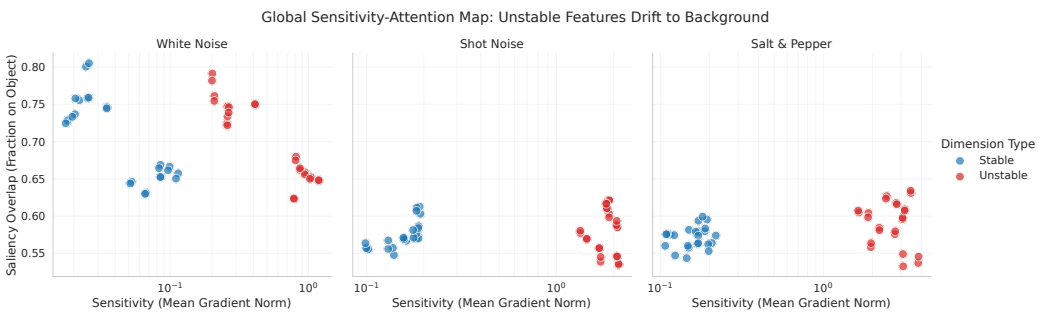


Figure E.1: **Decoupling brittleness from importance.** Each point represents a feature dimension. The horizontal axis is empirical sensitivity $d_{c,j}$, and the vertical axis is the reported attribution or saliency statistic. High-attribution coordinates can be stable or unstable; NCSP targets high-sensitivity coordinates.

Table E.2: BAR accuracy for absolute-deviation and squared-difference sensitivity variants.

Sensitivity score	BAR accuracy	Comment
Absolute difference	89.64 ± 1.02	Main method; estimates Eq. (2)
Squared difference	87.31 ± 1.44	More sensitive to rare large deviations

groups. The high-sensitivity group has more high-frequency energy in the analyzed BAR feature maps. This is an empirical observation for these datasets, not a required assumption of NCSP.

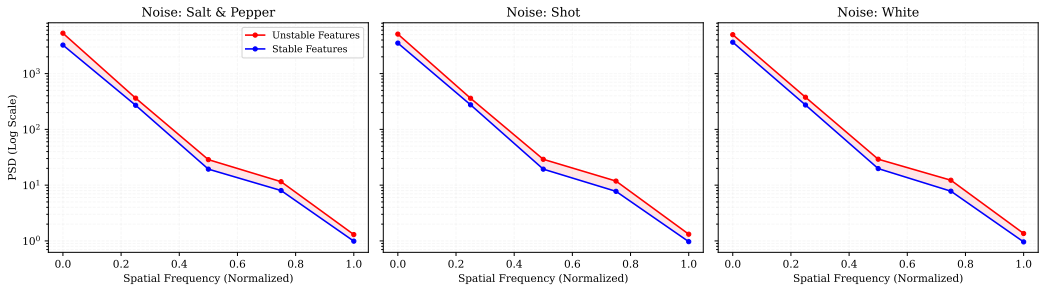


Figure E.2: **Power spectral density of stable versus unstable features.** Unstable features show heavier high-frequency tails in the analyzed BAR feature maps. Frequency is one mechanism for the observed sensitivity gap, not a universal requirement.

E.5 L1 ABSOLUTE DEVIATION VERSUS SQUARED-DIFFERENCE SENSITIVITY

The sensitivity estimand in Eq. (2) is an expected absolute scalar deviation:

$$S_{c,j} = \mathbb{E}\left[\left|\phi_j(T_\sigma(X)) - \phi_j(X)\right| \,\middle|\, Y = c\right].$$

The empirical estimator therefore uses elementwise absolute differences. Lemma 4.1 treats this absolute-deviation estimand as a finite-noise perturbation modulus; in smooth or stable-gradient regions, it reduces to a quantity proportional to the local input-gradient norm. It does not imply that the estimator should use squared differences.

A squared-difference variant is possible, but it overemphasizes rare large deviations and was less stable in our BAR experiments. Since NCSP only needs a reliable coordinate ranking, not a high-variance estimate of perturbation energy, the absolute-deviation estimator is the natural implementation of the theory.

Table F.1: Additional architecture results for NCSP. Main-paper comparisons use ResNet-18; larger-backbone results are diagnostic and should be reported with seed statistics when available.

Backbone	Waterbirds Avg.	Waterbirds WGA	BAR Avg.
ResNet-18	97.11	79.43	89.64
ResNet-50	95.32	90.03	91.44
ConvNeXt	96.00	86.83	96.94
ViT-B/16	93.86	83.04	93.27

F Additional Experimental Results

F.1 Sensitivity Estimation Budget

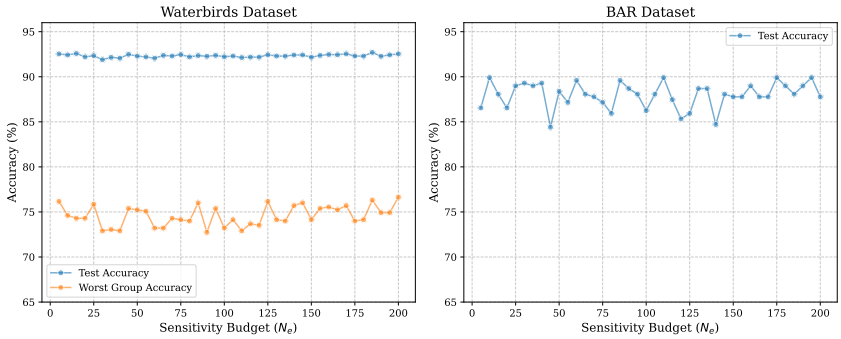


Figure F.1: Test accuracy and, where applicable, WGA as a function of the per-class sensitivity budget N_e on Waterbirds and BAR. The plot varies estimation budget while holding the pruning ratio fixed; it does not measure tolerance to arbitrary pruning.

Performance stabilizes with modest estimation budgets. This is consistent with Proposition 4.3: exact recovery requires a top-k gap, while near-optimal retained sensitivity depends on the uniform estimation error.

F.2 Multi-Backbone Results

The main tables use ResNet-18 for fair comparison with prior baselines. To test whether the sensitivity signal is specific to ResNet-18, we also evaluate NCSP on ResNet-50, ConvNeXt, and ViT-B/16. We treat these results as architecture-portability diagnostics rather than direct comparisons with baselines that report only ResNet-18.

These results support the claim that class-conditional sensitivity pruning is not tied to one convolutional backbone. At the same time, architecture differences matter: ViT-B/16 achieves strong average accuracy but lower Waterbirds WGA than ResNet-50 and ConvNeXt under the same simple protocol. We therefore treat architecture generality as evidence of portability, not as a claim that the same hyperparameters are optimal for every representation.

F.3 Per-Class and Per-Group Breakdowns

Table F.2 reports BAR per-class accuracy after NCSP. The gains are spread across classes; every class is well above the 60.51% ERM average reported in the main paper. The lowest class is Throwing, suggesting that this action may retain more residual shortcut reliance or have less stable semantic headroom.

For Waterbirds, we report worst-group accuracy in the main and multi-backbone tables.

Table F.2: BAR per-class accuracy for NCSP with ResNet-18.

BAR class	Accuracy
Climbing	90.00
Diving	94.03
Fishing	85.00
Racing	92.86
Throwing	75.00
Vaulting	87.88
Average	89.64

Table F.3: MetaShift group accuracy for NCSP. Majority/minority labels refer to the training correlation.

Group	Accuracy
Cat / Indoor	97.73
Cat / Outdoor	86.05
Dog / Outdoor	95.35
Dog / Indoor	86.05
Worst group	86.05

Table F.4: Reversed Waterbirds stress test. The target is background and the nuisance is bird type. ResNet18 (R18) with white/Gaussian noise is reported as mean $\pm$ standard deviation over three seeds.

Backbone	White WGA	Shot WGA	S&P WGA	White Avg.	Shot Avg.	S&P Avg.
ResNet-18	91.75 ± 0.32	91.71	91.66	95.10 ± 0.18	95.08	95.06
ResNet-50	92.86	93.22	92.95	95.36	95.46	95.32

F.4 Role-reversal stress test on Waterbirds

We test whether NCSP only prunes background-like features by reversing the Waterbirds prediction task: background becomes the target label, and bird type becomes the nuisance attribute. If the method were hard-coded to remove background or high-frequency texture, it should fail in this setting. Instead, it continues to identify the brittle nuisance direction.

This result supports the intended interpretation: NCSP tracks class-conditional brittleness in the learned representation, not a fixed semantic category such as "background." The role-reversal experiment also clarifies the frequency discussion in Appendix A.8: frequency is one mechanism that can produce sensitivity, but the operational diagnostic is measured perturbation response.

F.5 Adaptive pruning results

The adaptive pruning rule in Appendix B.3 reduces dependence on a fixed pruning ratio. We evaluate a global scaling factor $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$. Performance varies within a narrow range, suggesting that NCSP is not balanced on a sharply tuned pruning threshold.

Table F.5: Adaptive pruning sensitivity to the global scaling factor α. The rule sets $\tau_c = \mathrm{clip}(\alpha\,\mathrm{CV}_c, \tau_{\min}, \tau_{\max})$.

Scaling range	BAR accuracy range	Waterbirds WGA range
$\alpha \in \{0.5, 1.0, 1.5, 2.0\}$	88.4–90.1	78.0–80.9

The fixed $\tau = 0.25$ setting remains the main method because it is simple and directly comparable across experiments. The adaptive variant is useful when deploying NCSP on a new representation where the number of brittle coordinates is unknown.

F.6 CelebA as a scope-boundary experiment

CelebA is a difficult case for NCSP. The gender–hair-color shortcut may be encoded in intermediate representations rather than as a separable brittle final-layer subspace. Consistent with the scope conditions in Section 6, NCSP preserves high average accuracy but fails to substantially improve the Blond-Male worst group.

This negative result marks the boundary of the method. NCSP is a lightweight post-hoc intervention for settings in which shortcut directions are head-accessible and sensitivity-separated. When the

Table F.6: CelebA scope-boundary result for NCSP. Target is blond hair and nuisance attribute is gender. We include it as a failure case.

Group	Accuracy
Overall	94.88
Non-blond / Female	92.89
Non-blond / Male	98.90
Blond / Female	93.55
Blond / Male	52.22
Worst group	52.22

shortcut is deeply entangled in the backbone, methods that retrain representations, infer environments, or explicitly optimize hard groups may be preferable.

G Additional controls and stress tests

These controls isolate the sources of improvement and test the method outside its favorable cases. Clean-data head refitting accounts for much of the gain over ERM, sensitivity ranking does not uniformly beat matched masks, and label-only selection does not guarantee WGA. Unless a table states otherwise, the compared masked systems use the same data, pruning ratio, initialization, optimizer, regularization, and refit budget.

G.1 Component attribution

Table G.1: Controlled component ablation. BAR reports average accuracy; the other cells report average accuracy/WGA.

System	BAR	Waterbirds	MetaShift
Original ERM	60.51	90.10/63.74	83.30/69.60
Clean-data head refit only	91.93	98.30/74.60	91.45/86.98
Sensitivity mask only, no refit	68.99	91.83/40.49	88.21/68.16
Random equal-sparsity mask + refit	90.71	97.58/81.72	91.89/86.78
$\lvert W\rvert\times$sensitivity mask + refit	92.08	97.39/78.42	91.91/86.22
NCSP	89.64	97.11/79.43	92.49/86.05

Relative to refit only, NCSP improves Waterbirds WGA by 4.83 points and MetaShift average accuracy by 1.04 points, but it is worse on BAR, Waterbirds average accuracy, and MetaShift WGA. Random masking also matches or exceeds NCSP on several metrics. The main-table gains therefore belong to the complete pipeline; this ablation does not establish a uniform advantage for sensitivity ranking.

G.2 Label-only selection under latent-group skew

For a configuration θ, label-only validation accuracy is $A_{\mathrm{val}}(\theta) = \sum_g \pi_g A_g(\theta)$. A small π_g can hide a large error for group g, so average validation accuracy cannot guarantee WGA.

Across seven candidate selection runs, no run selected the no-intervention option; the mean selected-versus-oracle WGA regret was 5.94 points on MetaShift, 6.31 on Waterbirds, and 0 on BAR, where the result is based on two runs. These quantities are screening signals, not predictors of group performance. If representative group labels or shifted validation data are available, they should determine model selection.

Table G.2: Label-only selection compared with a WGA oracle. Oracle regret is the difference from the best WGA in the evaluated configuration set.

Dataset	Selection rule	Uses groups?	Test Avg.	Test WGA	Oracle regret
Waterbirds	Original ERM	No	90.10	63.74	–
Waterbirds	Fixed NCSP	No	97.11	79.43	2.49
Waterbirds	Label-only selected	No	97.23	79.12	2.80
Waterbirds	WGA oracle	Yes	–	81.92	0
MetaShift	Original ERM	No	83.30	69.60	–
MetaShift	Fixed NCSP	No	92.49	86.05	5.51
MetaShift	Label-only selected	No	–	85.63	5.93
MetaShift	WGA oracle	Yes	–	91.56	0

Table G.3: Pruning-ratio sweep over the reported moderate and high-pruning settings. Dashes denote combinations that were not measured.

Dataset / metric	$\tau = .05$	$\tau = .10$	$\tau = .15$	$\tau = .25$	$\tau = .40$	$\tau = .90$
BAR average	89.79	89.64	89.79	89.64	90.71	80.92
Waterbirds average	96.19	96.82	95.92	97.11	96.68	–
Waterbirds WGA	75.30	73.67	77.56	79.43	77.09	61.99
MetaShift average	92.03	92.15	91.80	92.49	92.03	–
MetaShift WGA	86.44	84.97	83.52	86.05	86.06	–

G.3 Pruning ratio and aggregation

The moderate range is not sharply peaked, so $\tau = 0.25$ is a fixed baseline, not an estimate of the nuisance dimension or a universal optimum. The reported $\tau = 0.90$ cells show that aggressive pruning can remove useful evidence.

Table G.4: Robust aggregation ablation. Mask overlap is measured against the mean-based mask.

Dataset	Aggregator	Accuracy	WGA	Mask overlap
MetaShift	Mean	92.49	86.05	1.000
MetaShift	Median	92.84	87.58	0.728
MetaShift	10% trimmed mean	92.03	87.55	0.853
Waterbirds	Mean	97.11	79.43	1.000
Waterbirds	Median	97.07	77.80	0.785
BAR	Mean	89.64	–	1.000
BAR	Median	92.24	–	0.817

The alternative aggregators change the selected masks but remain competitive. The median improves MetaShift and BAR but lowers Waterbirds WGA, so no aggregator is uniformly preferable and it must not be chosen using test-group metrics.

G.4 Coordinate basis and many-class stress tests

Table G.5: Orthogonal-basis stress test. For each orthogonal Q, replacing h by Qh and W by $WQ^\top$ preserves the original predictor exactly.

Dataset	Pre-NCSP agreement	Native Avg./WGA	Rotated Avg./WGA
MetaShift	15/15	92.49/86.05	91.70/87.94
Waterbirds	10/10	97.11/79.43	97.27/77.13
BAR	10/10	89.64/–	91.66/–

Post-intervention differences arise solely from the coordinate basis. NCSP can exploit structure in native learned coordinates, but it does not identify an intrinsic nuisance subspace. A preliminary MetaShift alignment analysis also gave $77.5 \pm 17.7\%$ purity, sensitivity Gini 0.182, and a negative dissociation of -13.23 ± 0.21 points; unlike the BAR and Waterbirds entries in Table E.1, this used a different probe and remains preliminary.

Table G.6: Joint class–context WGA on an unofficial 20-class, six-context NICO++ fallback (Zhang et al., 2023). Each row is a single run.

Method	ResNet-50	ViT-B/16
ERM	50.00	50.00
Refit only	60.00	50.00
Random mask + refit	66.67	50.00
Magnitude mask + refit	66.67	50.00
NCSP	65.79	50.00

The many-class experiment does not show a sensitivity-specific advantage: equal-sparsity controls match or exceed NCSP. Because this fallback is not the official 60-class Track-1 release and has one run per backbone, it marks an empirical boundary rather than settling large-class scaling.

G.5 Practical interpretation of the diagnostics

The ablations suggest a concrete workflow for applying NCSP beyond the datasets studied here. First, train or obtain a standard ERM model and freeze its representation. Second, estimate d_c using a modest per-class budget and a mild probe. Third, compare top-sensitive pruning against random or bottom-sensitive pruning on label-only validation data. If top-sensitive pruning preserves clean accuracy while bottom-sensitive pruning degrades it, the representation likely contains a removable brittle subspace. Finally, perform the constrained head refit and evaluate the resulting classifier on any available distribution-shift or group metric. This workflow uses group labels only for final reporting when such labels exist.

The diagnostic can falsify its own premise. If sensitive coordinates are not removable, then pruning them is likely to harm semantic prediction. If prediction stability does not improve after masking, then the selected coordinates may not control the relevant logits. If class-conditional and global masks behave similarly, then shortcut information may be shared across labels rather than class-specific. Treat these outcomes as evidence about the representation rather than as hyperparameter failures. In this sense, NCSP is a representation audit as much as a debiasing intervention.

The constrained refit is essential to the evaluated pipeline: masking without refitting performs poorly in Table G.1. At the same time, refit-only and random-mask controls explain much of the ERM-to-NCSP gain. Once a mask is fixed, the recovery problem is convex and cannot relearn a masked connection, but the experiments do not show that sensitivity ranking is uniformly the best way to choose that mask.

These diagnostics also clarify how to compare NCSP with stronger baselines. Methods such as JTT, DEDIER, or PDE may obtain higher worst-group accuracy on particular datasets because they more directly optimize hard examples or group-robust behavior. NCSP should therefore be evaluated as a lightweight post-hoc method with a small computational and annotation footprint. It does not dominate every worst-group benchmark. Its main advantage is that it exposes and removes a specific, interpretable failure mode of the final head when the stability–semantics gap is present.

H SCOPE, FAILURE MODES, AND RESPONSIBLE USE

NCSP can fail or give small gains in the following cases.

- **Stable spurious cues.** If the nuisance cue does not change under the probe, sensitivity cannot separate it from semantic features, and NCSP may leave the shortcut in place. Check the reverse-pruning and probe-dissociation diagnostics before using the intervention.
- **Sensitive semantic cues.** Fine-grained tasks can depend on texture, material, or local edges that carry real class information. Aggressive pruning removes this evidence and lowers accuracy. High sensitivity alone does not mark a feature as spurious.
- **Shortcuts outside the head, or not axis-aligned.** A head-only mask cannot remove a shortcut encoded through nonlinear interactions in the backbone, or one spread across many coordinates. This is the likely cause of the CelebA failure (Appendix F.6).
- **Weak frozen representations.** NCSP assumes the frozen representation already contains stable semantic coordinates. If ERM did not learn them, refitting the head cannot recover them.
- **Augmentation-trained backbones.** If the backbone was trained with augmentations close to the probe, such as Gaussian noise or color jitter, the sensitivity gap shrinks and NCSP gains less. The model has already suppressed the instability the probe measures.
- **Nonlinear heads and intermediate layers.** The convexity guarantee needs a frozen representation and a linear head. The method does not cover nonlinear heads, intermediate-layer pruning, or input-space interventions. Local linearization of nonlinear heads and structured pruning of intermediate layers are separate directions for future work.
- **Per-sample masks.** A test-time mask for each input would need extra forward passes at inference and would be harder to reproduce. Class-conditional masks are more specific than one global mask and still keep sensitivity estimation offline.
- **Text benchmarks.** Text datasets such as CivilComments have different representation geometry and perturbation semantics. Applying NCSP to them would need a new definition of label-preserving perturbations and coordinate sensitivity. We do not study text here.
- **Fairness.** Higher worst-group accuracy on benchmarks is not a fairness certificate. NCSP prunes and refits the classifier head on a frozen visual representation. It does not prove causal invariance or remove all sources of bias. Deployed systems still need group-aware evaluation whenever group labels are available.