WHEN SHORTCUTS RETURN: POST-RESTORATION ROBUSTNESS IN GRAPH CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporarily suppressing a known shortcut can improve robustness to distribution shift, but an endpoint evaluation does not reveal how that robustness changes once the shortcut returns. We study this question on a controlled graph-classification task by permuting node colors during a training prefix and then restoring the original inputs. Permanent permutation achieves 0.995 worst-environment accuracy in a matched comparison, versus 0.983 for a released prefix, making it an essential accuracy baseline. A trajectory experiment with ten independent data seeds and full checkpoint forks then measures the restoration horizon directly: every released run starts above 0.99 at restoration, and 36 of 40 meet a sustained-degradation criterion within 6,000 restored-input updates, with continuous suppression exceeding the released arms by 0.55–0.76 anti-environment accuracy (seed-paired intervals excluding zero). Batch size modulates the re-acquisition at matched update count, and a longer suppression prefix helps only at the larger batch. Optimizer-state forks with update-norm-matched controls show the reset damage is largely a scale-shock artifact: a shock-only branch reproduces most of the loss while norm matching does not rescue it. Separate fixed-data controls show that environment grouping changes the effect of an invariance penalty, while diverse training mixtures can support robust ERM without suppression. An external task with naturally diverse shortcut proportions reverses the ordering—plain ERM is robust and continuous suppression is counterproductive—bounding when suppression is the right tool. We distinguish these observations from unresolved mechanisms and provide conditional analysis for low CE gradients and fixed-feature replacement. The results motivate evaluating both suppression and restoration horizons when studying the persistence of robustness.

1 INTRODUCTION

Removing an informative shortcut during training can make a predictor use other features. A less settled question is what happens when that shortcut returns while training continues. A high accuracy immediately after restoration can reflect a durable solution or a temporary dependence on training history. An endpoint evaluation alone cannot distinguish them. We study this distinction in attributed graph classification, where a known color block provides a controllable shortcut and graph motifs define the labels.

Our intervention is deliberately privileged: it requires knowing the shortcut columns. We permute colors across nodes of different training graphs for a finite prefix, then restore the original inputs. The aim is to measure the persistence of the resulting robustness, rather than to outperform deletion of a known nuisance feature. Indeed, under a matched full-label comparison, permanent training-time permutation achieves 0.995 ± 0.001 worst-environment accuracy, compared with 0.983 ± 0.006 after a finite permutation prefix. Keeping the shortcut suppressed is a stronger accuracy baseline in this setting.

The restored-input continuation provides the more informative comparison. On a short, nominal 1,440-update schedule, batch sizes 64 and 256 give anti-environment accuracies of 0.824 ± 0.075 and 0.824 ± 0.046. At 5,640 total updates, both perform poorly (0.141 ± 0.034 and 0.083 ± 0.006).

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 dedicated trajectory experiment with ten independent data seeds and full checkpoint forks (P1) then measures the restoration horizon directly: every released run starts above 0.99 at $H{=}0$, 36 of 40 degrade within $H{=}6{,}000$ restored updates, and continuous suppression exceeds the released arms by $+0.55/{+}0.76$ (batch 64/256, seed-paired intervals excluding zero). Batch size remains operative at matched update count (0.442 vs. 0.239 final at $T_w{=}200$), and a longer prefix helps only at the larger batch. These observations establish horizon dependence with independent-seed replication; they do not identify the re-acquisition rate or its single cause.

Two further comparisons delimit this result. Norm-matched optimizer forks show the reset damage is largely a scale-shock artifact, not evidence for a particular memory mechanism. Separately, a shortcut-diverse training mixture permits strong ERM generalization without suppression, and changing V-REx's environment partition on fixed graphs substantially changes its effect. Input history, training distribution, and grouping are distinct interventions and receive distinct explanations.

The contribution is a controlled empirical account: (i) continuous suppression, finite suppression, and the post-restoration continuation as distinct objects; (ii) schedule controls showing a released model's robustness can disappear over a longer horizon; (iii) optimizer and distribution controls separating measured sensitivities from causal claims, plus an external task bounding when suppression applies. Conditional bounds for low CE gradients and fixed-feature replacement are supporting analysis, not a theory of GNN feature selection.

2 Experimental Setting and Evidence Conventions

Graphs and shifts. RedHouse contains three motif classes (house, cycle, crane) attached to background graphs from BA, ER, and WS families. The label is the motif class. The background family matches the class with probability b_{struct}; node colors are sampled with

$$P(C = Y) = \rho, \qquad P(C = c \neq Y) = \frac{1-\rho}{K-1}, \quad K = 3.$$

The two generation parameters can be varied separately. The recorded evaluations use structure-shift, feature-shift, both-shift, and anti-correlated environments. The audited generator defaults are $(b_{\mathrm{struct}}, \rho) = (1/3, 0.8)$, $(0.8, 1/3)$, $(1/3, 1/3)$, and $(0.2, 0.2)$, respectively. The first two retain 0.8 on the other axis, even when training uses 0.9; thus these are not exactly single-axis shifts from the training law. Motifs and their labels remain intact. Dataset hashes are needed to identify individual realizations.

What constitutes a shortcut here. For this color rule, independence is at $\rho = 1/3$. Values below $1/3$ create a reversed association, rather than progressively less information. At $\rho = 0.1$, a color-only logistic audit reaches 97.5% accuracy; at $\rho = 0.9$ it reaches 100%, and at $1/3$ it reaches 36.8%. The graph-level color histogram can be much more predictive than a single node. We therefore call ρ a match probability, and do not identify $1 - \rho$ with a graph-level error floor.

Model and intervention. The principal encoder is a three-layer GIN with hidden dimension 64 and sum pooling, trained using Adam with learning rate 10^{-3} and coupled weight decay 10^{-4} in the audited experiments. The encoder sums its three layer outputs before a projection and graph pooling; dropout defaults to zero and it contains no batch normalization. Node inputs contain degree features and color one-hots. During suppression, one fresh uniform permutation exchanges color rows across the concatenated nodes of the training batch, preserving its overall color histogram. At the release step T_w, the original inputs return. The graph topology, labels, and non-color input columns remain available throughout. Suppression requires the known color-column indices; this information advantage is explicit in all method comparisons.

Training clocks. We distinguish suppressed updates T_w, restored-input updates H, and total $T = T_w + H$; a run ending before release is continuous suppression. Batch size changes both updates per epoch and examples per update—matching epochs, updates, and examples are different comparisons, never silently substituted.

Protocols and metrics. We report each protocol locally rather than pooling results across them. The full-label comparison uses 3,000 graphs, batch 256, and 60 epochs; the historical low-label table

Table 1: Full-label comparison (E52/E53): 3,000 training graphs, batch 256, 60 epochs, three seeds. Reported worst-environment accuracy, mean ± standard deviation. Permanent permutation and finite-prefix permutation both use original test inputs. The zeroing arm zeros colors in both training and testing, as verified in the evaluation code.

Training intervention	Known color block	Release	Accuracy
ERM	no	not applicable	0.069 ± 0.007
Permanent permutation	yes	never	0.995 ± 0.001
Finite-prefix permutation	yes	$T_w = 200$	0.983 ± 0.006
Train-and-test zeroing	yes	never	0.994 ± 0.001

uses 300 labeled graphs and approximately 120 updates. The latter never releases a $T_w = 200$ mask. Schedule studies use nominal 1,440- or 5,640-update budgets, with a separate 1,410-update rerun explicitly distinguished in App. A. Worst-environment accuracy is the minimum over the evaluated environments; anti-environment accuracy refers to one particular environment. Where available, uncertainty is the reported population-form standard deviation across seeds (`ddof = 0`), not a confidence interval. E41, E44-b, and E40 vary optimization seeds on a single dataset realization; E52 and E62 also vary the generated dataset with the seed. Results with different metrics, seeds, or budgets are not treated as paired estimates. Experiment identifiers link observations to their source records. The P0–P2 and P4 rewrite experiments run on CPU (deterministic, four threads); the historical E40–E65 logs were produced on GPU, and the P0 anchor audit found release-sensitive arms to be environment-sensitive, so cross-environment endpoints are not pooled.

Limits of the graph task. A no-message-passing feature aggregate reaches 0.876 worst-environment accuracy when colors are zeroed and 0.08 with colors present (E63). Degree histograms alone reach 0.64. Thus structural statistics provide substantial label information; these experiments do not establish that message passing is necessary for the observed competition.

3 Suppression Helps, but Restoration Changes the Question

Matched accuracy comparison. Table 1 contrasts permanent suppression and a finite prefix using the same full-label experiment family. The reported advantage of permanent permutation is 0.012 in mean worst-environment accuracy. Recomputation from the three per-seed records gives paired gains of 0.0073, 0.0067, and 0.0207, with mean 0.0116. Three pairs are insufficient for a precise population uncertainty estimate. The comparison nevertheless makes the practical baseline clear: if the color block is known to be unnecessary at deployment, permanent suppression is competitive and simpler than choosing a release time.

The release clock must be audited. The E52 implementation makes 12 updates per epoch, hence 720 in total: the finite-prefix arm has $T_w = 200$ and $H = 520$. A separate duration sweep uses the same nominal 60-epoch budget but selects the best in-distribution validation checkpoint. Its $T_w = 800$ and 1600 arms never release, and even a shorter-prefix arm's selected checkpoint may precede release. Therefore its approximately 98.5–98.6% scores cannot establish a common post-restoration plateau or a universal 100-step threshold. We use the final-checkpoint E52 comparison here.

Later continuation can undo part of the gain. A trajectory with $T_w = 200$ and 1,240 restored-input updates ends at 0.836 anti-environment accuracy, versus 0.998 under permanent permutation (App. C). The dedicated restoration-horizon experiment (P1, Sec. 4; ten independent data seeds, full checkpoint forks) confirms and extends this: every released run starts above 0.99 at $H{=}0$, and 36/40 degrade by the sustained criterion within $H{=}6{,}000$, with the continuous-permutation contrast at $+0.55/{+}0.76$ (batch 64/256). The finite-prefix intervention therefore demonstrates persistence over particular horizons, not permanent removal of shortcut dependence, and the re-acquisition onset varies widely across seeds.

Table 2: Restoration-horizon experiment (P1): ten data seeds, mean ± sample SD. AAUC: time-averaged anti accuracy over $H \in [0, 6000]$. "deg.": sustained drop $\geq$10 points below $H{=}0$ over three consecutive evaluations (all runs start above 0.99 at $H{=}0$). Paired differences (continuous − released) with seed-paired bootstrap 95% intervals.

Arm / batch	final anti	AAUC	deg.	paired diff. vs. continuous
ERM / 64	0.084 ± 0.044	–	–	–
ERM / 256	0.061 ± 0.013	–	–	–
Continuous / 64	0.996 ± 0.002	–	–	–
Continuous / 256	0.996 ± 0.003	–	–	–
Release $T_w{=}200$ / 64	0.442 ± 0.230	0.776	9/10	$+0.554$ [0.407, 0.680]
Release $T_w{=}200$ / 256	0.239 ± 0.211	0.572	10/10	$+0.757$ [0.629, 0.875]
Release $T_w{=}800$ / 64	0.532 ± 0.298	0.746	8/10	$+0.465$ [0.288, 0.641]
Release $T_w{=}800$ / 256	0.514 ± 0.225	0.610	9/10	$+0.482$ [0.349, 0.609]

Continuous suppression is a different result. Under the 10%-label protocol, $T_w = 200$ exceeds the approximately 120-update budget. Its 89.7% worst-environment accuracy and the 92.9% result with consistency regularization concern continuous permutation, not restoration. We retain these results as supplementary accuracy comparisons (App. D) and do not use them to support claims about a finite prefix. The audited adaptive controller likewise never releases within its tested budget. Its high accuracy is evidence about sustained suppression, not validated schedule discovery.

4 Post-Restoration Exposure Changes the Batch-Size Interpretation

An epoch-fixed batch sweep appears to show strong batch-size dependence: with $T_w = 200$, anti-environment accuracy is 0.113 at batch 64 and 0.937 at batch 512 in the initial run. Larger batches, however, perform fewer updates over the same epochs. With 3,000 training graphs, the reported 120-epoch configurations perform approximately 5,640 and 720 updates, respectively. Their restored-input horizons are consequently very different.

Trajectory experiment (P1). We then measured the restoration horizon directly: ten independent data seeds (with paired model, data-order, and permutation seeds), batches 64 and 256, four arms each—ERM, continuous permutation, release at $T_w{=}200$, and release at $T_w{=}800$ with full checkpoint forks (optimizer state, data-stream RNG, and cursor inherited; $H{=}0$ evaluated before the first restored update). Figure 1 shows the anti-environment trajectories; Table 2 reports the terminal and integrated statistics.

Three readings. *(i) Degradation after release is the norm:* all forty released trajectories start above 0.99 at $H{=}0$ and 36/40 meet the sustained-degradation criterion within $H{=}6{,}000$; the continuous-permutation contrast is $+0.55/+0.76$ (batch 64/256), intervals excluding zero. *(ii) Batch size remains operative at matched H:* batch-256 releases degrade more than batch-64 (0.239 vs. 0.442 final at $T_w{=}200$)—update count is first-order, batch size modulates re-acquisition at the same count. *(iii) A longer prefix helps only at batch 256:* $+0.275$ [0.114, 0.431] at 256, unresolved at batch 64 ($+0.089$ [-0.150, 0.310]); the schedules are not pooled.

What has and has not been isolated. These controls establish horizon-dependent robustness with independent data seeds; they do not separate biased-input updates from examples seen, parameter displacement, or the LR schedule, and degradation onset varies widely across seeds. The microbatch-accumulation cell remains excluded (92 of 1,440 updates under-filled; App. A); the pure-batch arms are unaffected.

Suppression strength is another variable. On the short E48-b schedule, free permutation, histogram-matching, and cross-class rotation reach $0.831/0.071/0.556$; their color-histogram probes are $0.35/1.00/0.87$ (E59). Histogram matching is an identity control (permutes identical rows within class–color groups); rotation mostly renames the association. These support measuring the residual shortcut after corruption—not a dose-response curve from three operations.

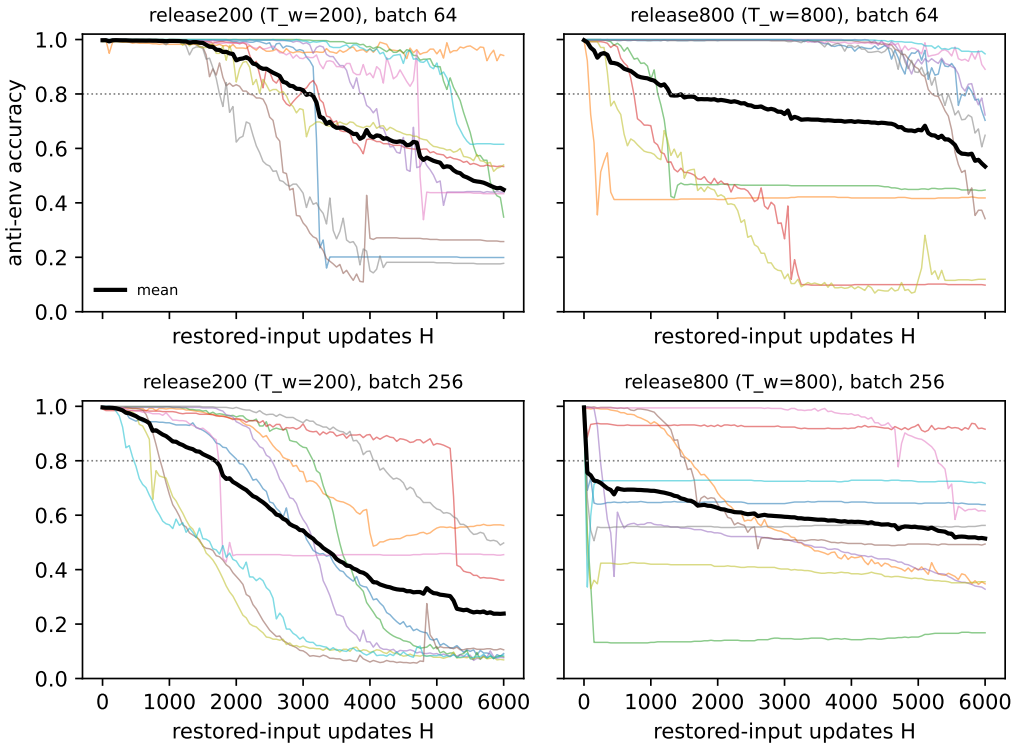

Figure 1: Anti-environment accuracy vs. restored-input updates H (thin: individual seeds; thick: mean). Every released run starts above 0.99 at $H{=}0$; degradation thereafter is the norm (36/40 runs meet the sustained-drop criterion), with large seed-to-seed variation in onset and severity.

5 Optimizer Forks Establish Sensitivity, Not Its Cause

We distinguish two starting points: a model immediately after a suppressed prefix, and a model trained on original inputs until low training loss. The same state reset need not have the same meaning at these points.

Forking at restoration. E40 forks the $T_w = 200$ checkpoint and continues each branch for 1,000 updates with the same data order. The four-seed final anti-environment accuracies are reported in Table 3. Resetting only the first moment changes little in these runs; resetting the second moment or all state produces larger losses. These contrasts establish intervention sensitivity at this checkpoint.

Norm-matched forks separate shock from state removal (P2). A dedicated fork experiment (P2; ten independent data seeds, batch 256, $T_w{=}200$ prefix re-run under the P1 stream protocol, six branches per seed, $H{=}1{,}000$) tests whether the second-moment reset's damage is an update-magnitude effect. Branches: F0 unchanged; F1 $v{=}0$ only (the E40 intervention); F2 clear $m/v/t$; F3 v-reset with every step's global update norm rescaled to F0's same-step norm; F4 the same with per-layer norms matched; F5 no state clear but every step's norm rescaled to F1's (shock-only). Final anti-environment accuracy: F0 $0.913{\pm}0.064$; F1 $0.197{\pm}0.065$; F2 $0.646{\pm}0.115$; F3 $0.222{\pm}0.109$; F4 $0.242{\pm}0.121$; F5 $0.269{\pm}0.072$ (seed-paired bootstrap intervals in App. C). Three readings. *(i) Shock alone reproduces most of the damage:* F5 removes no state yet loses 0.644 [0.593, 0.692] vs. F0—nearly the full F0–F1 gap of 0.716 [0.669, 0.760]. *(ii) Norm matching does not rescue:* F3–F1 $= +0.025$ $[-0.040, 0.087]$—within noise; the early steps or a non-magnitude channel do the damage. *(iii) Stale m with a zeroed denominator is the worst combination:* full clearing (F2) is 0.448 [0.383, 0.528] better than F1 and within 0.268 [0.186, 0.344] of F0. The v-only reset is thus a scale-shock artifact of the retained first moment, not evidence that the second moment carries the warmup robustness.

Table 3: Restoration-point forks (E40): anti-environment accuracy for each of four seeds after 1,000 continuation updates. m, v, and t denote Adam's first moment, second moment, and step counter. The evaluation uses the first 512 graphs of each test environment; all four optimization seeds share dataset seed 0.

State intervention	Seed 0	Seed 1	Seed 2	Seed 3
Unchanged	.936	.941	.850	.920
Reset all	.578	.465	.414	.355
Reset m only	.891	.938	.850	.916
Reset v only	.107	.207	.156	.115
Reset t only	.951	.945	.836	.914

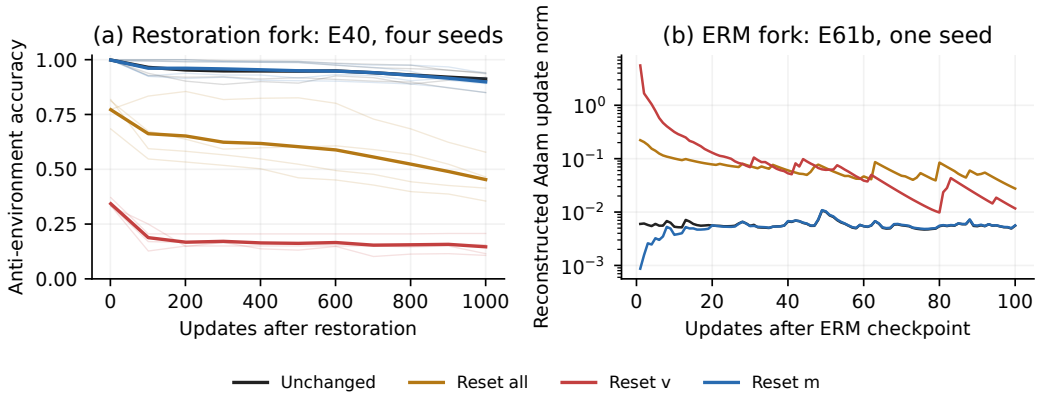

Figure 2: Two different fork experiments, replotted from the saved records. (a) E40 starts after 200 suppressed updates; thin curves show four optimization seeds, thick curves their means on a fixed dataset. (b) E61b starts after 400 ordinary ERM updates and uses one seed. Norms are reconstructed from Adam state, not stored parameter differences. The right panel does not measure update scales for the left panel.

Update scale is a major confound. The additional E61b audit reports initial reconstructed update norms of 0.32–5.6 for a second-moment reset compared with approximately 0.006 for the unchanged branch, a peak ratio near 900. The source code confirms that this measurement starts at the step-400 *ERM* checkpoint, not E40's restoration checkpoint. The maximum matched-step ratio recomputed from the rounded log is approximately 930. It is evidence of a reset-induced scale shock at the ERM checkpoint, not a measurement of the shock magnitude at restoration. Regardless, preserving m while resetting the denominator state changes update scale as well as preconditioning. A resulting accuracy loss cannot by itself identify which property maintained robustness. A step-counter-only control also cannot rule out interactions among state components. The P2 experiment above is the direct test at the restoration point; this audit remains the ERM-point evidence.

Forking after ordinary ERM fitting. At the E61 checkpoint (step 400, training loss 3×10^{-5}), 1,000-step continuations give 0.088 ± 0.010 for unchanged state, 0.091 ± 0.017 for a full reset, and 0.111 ± 0.018 for a second-moment reset. The tested resets do not repair this model on the same biased objective. This does not imply that its structural information is absent, or that all alternative optimization paths would fail.

Scope of the optimizer evidence. An additional long-schedule comparison reports approximately 0.75 anti-environment accuracy for SGD with momentum and 0.11 for Adam at batch 64. Learning rates and optimizer dynamics differ, so this is a configuration comparison rather than an isolated second-moment effect. To test a stronger explanation, a reset must be compared with an update-norm-matched reset and a no-reset branch receiving matched update shocks. Those experiments have not been completed in the evidence used here.

Table 4: Fixed-data partition control (E62), three seeds. Reported worst-environment accuracy, mean ± standard deviation. All V-REx rows use $\beta = 100$. These results are not pooled with Table 1.

Objective / partition	Penalty start	Accuracy
ERM / none	not applicable	0.890 ± 0.056
V-REx / topology	0	0.365 ± 0.160
V-REx / random	0	0.884 ± 0.054
V-REx / ρ	0	0.929 ± 0.023
V-REx / ρ	100	0.946 ± 0.017

6 Training Mixtures and Environment Partitions Provide a Separate Route

Robustness can improve without temporarily hiding the shortcut. This distinction matters when interpreting the finite-prefix intervention: its benefit in one homogeneous training distribution does not imply that biased features must always be removed early.

Mixture comparison. Under E55's full-label protocol, a mixture with graph-level match probabilities $\{0.5, 0.7, 0.9\}$ supports ERM worst-environment accuracy of 0.874 at 60 epochs and 0.958 at 240 epochs. E58 compares that mixture with homogeneous $\rho = 0.7$, holding the mean match probability fixed. Homogeneous training gives 0.21 ± 0.05 and 0.42 ± 0.27, whereas the mixture gives 0.88 ± 0.05 and 0.95 ± 0.01 at the two budgets. Mean match probability therefore does not determine these outcomes. The mixture also changes the distribution of graph-level shortcut difficulty; these experiments do not isolate heterogeneity from the fraction of examples with a less reliable shortcut.

Partition comparison on fixed graphs. E62 keeps the diverse training graphs fixed and changes the environment tags used by V-REx. Table 4 includes ERM, which does not read those tags. The topology partition substantially reduces accuracy relative to ERM, whereas the random partition has a similar mean. The ρ partition gives a larger mean, but the per-seed gains over ERM are -0.0480, 0.0267, and 0.1393 (mean 0.0393). This heterogeneous three-pair result does not establish a precise or uniformly positive benefit. A 100-step delay also performs well; this particular comparison does not support a requirement to activate the penalty before the shortcut is fitted.

Why the partition result is not yet a mechanism. E62b finds approximately $0.9/0.05/0.05$ class proportions within topology groups, together with larger between-group degree-histogram differences. The ρ and random groups are class-balanced. Consequently, partition changes affect class composition and structural statistics as well as shortcut reliability. Similar majority-color-rule accuracies across groups do not establish equality of conditional risks. A class-balanced or class-conditional-risk topology control is needed to isolate the role of label composition.

E65 further reports a scaled V-REx penalty-gradient norm about 50 times the ERM gradient near step 10 in an unsuccessful topology-partition run. Thus that failure is not explained by an absent early penalty gradient. The gradient's magnitude alone does not identify its effect on the learned representation. We interpret the mixture and partition results as separate distributional interventions, rather than consequences of a universal optimization failure shared by all robust objectives.

External check: suppression is not universally the right tool (P4; App. C). An external MNIST-superpixel task (GOOD concept-shift protocol; five training environments with naturally diverse shortcut proportions 0.75–0.95—the mixture regime above) reverses the RedHouse ordering: across ten optimization seeds on the fixed split, plain ERM is robust (OOD 0.868 ± 0.003), *continuous* suppression is harmful (0.258 ± 0.021; paired ERM$-$suppression $= +0.610$ [0.598, 0.624]), and both released arms are indistinguishable from ERM (0.868 ± 0.003/0.869 ± 0.003; paired differences within ± 0.001). Suppression cannot help in this regime: the digit signal requires the corrupted channels (intensity-only inputs leave training CE at chance under the same architecture), and recolorization injects input-norm variance. The boundary condition: *suppression is necessary when*

a strong shortcut is homogeneous across training, and counterproductive when shortcut strength is naturally diverse—on both tasks the mixture, not the intervention, is the operative variable.

7 What the Supporting Analysis Can Explain

The mathematical analysis is deliberately separated from the empirical claims. Complete statements and proofs are in App. B.

Low CE gradients need both margins and Jacobians. For a parameter block B, a training margin lower bound $m_t \geq 0$, and a uniform logit-Jacobian bound M_t, the CE gradient satisfies $\|\nabla_B \widehat{L}\| \leq \sqrt{2}(K-1)M_t e^{-m_t}$. Summing ordinary GD updates bounds motion after a specified checkpoint. This calculation cannot identify the feature that produced the margin; nor does small CE imply small Adam updates. The experiments do not measure the required Jacobian envelope. First-layer gradient ratios are descriptive quantities under the current parameterization, not estimates of independent feature-learning rates.

Independent replacement in a fixed-feature comparison. In centered linear least squares, replacing a shortcut feature by an independent same-marginal copy decouples its coefficient; with zero initial shortcut coefficient, population GD keeps it zero during replacement. Freezing the structural predictor afterward, the restored linear component's L^2 norm is bounded by the structural residual uniformly over the restoration horizon. This model does not explain degradation from joint changes to a nonlinear representation.

Implemented permutation leaves different dependencies. For a batch of N nodes, a graph of size n_g has expected permuted color sum n_g times the batch mean, and the source pool includes its own colors: the batch operation is not independent graph-feature replacement, so these calculations are not guarantees for the GNN intervention.

8 Related Work

Shortcut learning and simplicity bias describe the preference for predictive cues that need not survive a distribution shift (Geirhos et al., 2020; Kalimeris et al., 2019; Shah et al., 2020). Gradient starvation studies how learning one feature can suppress others (Pezeshki et al., 2021); we study continued training after a known cue is restored, without proposing a new starvation theory. The distinction between low loss and limited parameter motion also matters in separable logistic regression (Soudry et al., 2018). IRM and V-REx use training environments to constrain predictive behavior (Arjovsky et al., 2019; Krueger et al., 2021); DIR, GSAT, and CIGA pursue related robustness goals on graphs (Wu et al., 2022; Miao et al., 2022; Chen et al., 2022), and GOOD provides a graph OOD benchmark framework (Gui et al., 2022). Our fixed-data partition comparison examines sensitivity to the grouping provided to one objective—not an impossibility result for these methods. Last-layer retraining can exploit retained representation information (Kirichenko et al., 2023); failed probes under shifted conditions here do not invalidate that premise.

9 Limitations and Conclusion

The evidence is strongest for a limited conclusion: finite suppression can produce high shifted-environment accuracy, yet that benefit depends on how long training continues after the shortcut returns; under fixed epochs, batch size changes this continuation budget and can obscure the dependence, so the two clocks are reported separately throughout.

Several questions remain open: no re-acquisition timescale is identified independently of other training variables; environment partitions change class composition as well as shortcut reliability; linear-probe failure under shift does not prove structural information was never learned; and the main task admits strong degree-feature predictions, with Cora profiling not validating the restoration result. In sum: permanent suppression is the right baseline when the nuisance block is known and the training shortcut is homogeneous; with diverse shortcut proportions, plain ERM is robust and suppression is counterproductive; restoration is a measurement instrument.

REFERENCES

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

Yongqiang Chen, Yatao Zhang, Yiwen Bian, Han Yang, Ma Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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*, 2:665–673, 2020.

Shurui Gui, Xiner Wang, Haoyu Zhang, Lu-Wei Jiang, Jiuxiang Feng, Carl Hayashi, Shaochun Liu, and Jiliang Tang. Good: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2022.

Dionysis Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (V-REx). In *International Conference on Machine Learning (ICML)*, 2021.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning (ICML)*, 2022.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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 (NeurIPS)*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Ying-Xin Wu, Xiao Wang, Aosong Zhang, Xiang He, Yuan Fang, Yu-Ye Liu, Jiuxiang Feng, and Jie Tang. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

AI USE STATEMENT

AI assistants were used in coding and in drafting and restructuring the manuscript and theoretical analysis. The authors are responsible for independently verifying the arguments, numerical results, and source attributions before submission.

REPRODUCIBILITY STATEMENT

The protocol inventory in App. A distinguishes the experiment families and unresolved record issues. Reproducing the experiments requires configurations and raw logs in addition to the manuscript source. The current rewrite adds no new training results.

A Protocol Inventory and Reproducibility Requirements

Evidence status. The manuscript reorganizes previously reported experiments. It does not add new training runs. The local result JSONs and corresponding training scripts were inspected, and the seed-level E52/E41/E44-b/E62 values were recomputed. A file-hash manifest accompanies the rewrite; code inspection is not proof that every historical run used the current code. Protocol details below retain the distinctions in those records; unavailable details are not inferred from rounded epoch counts. In particular, the exact release and last-update indices are needed before interpreting any endpoint as post-restoration evidence.

Table 5: Protocol families used in the main text. "Nominal" identifies a stated schedule, not verification of its realized update count.

Family	Budget and sample information	Interpretation
E52/E53	Full labels, 3,000 graphs, batch 256, 720 updates; 3 seeds	Matched suppression comparison; $T_w = 200$, $H = 520$
E41	5,640 total updates, $T_w = 200$; 5 seeds	5,440 restored-input updates
E44-b	1,440 updates, $T_w = 200$; 5 seeds	Pure-batch arms verified; accumulation arm excluded
E48-b	1,410 realized updates in the 3-seed rerun	Permutation-operation control
E40	Release at 200; 1,000 further updates; 4 seeds	Restoration-point state forks
E61	ERM checkpoint at 400; 1,000 further updates; 4 seeds	Fitted-ERM state forks
E61b	100-step update audit, seed 0	ERM checkpoint at 400, not restoration
E55/E58	Full labels, 60/240 epochs	Training-mixture comparison
E62/E62b	Fixed diverse graphs; 3 training seeds	Partition comparison and data audit
Low-label	300 labeled graphs, approximately 120 updates	$T_w = 200$ never releases

Audit corrections. The zeroing arm in E52 clears colors during both training and testing; permanent permutation tests on raw inputs. Its V-REx arms use different, shortcut-diverse training graphs, so they are not included as method-only contrasts in Table 1. The duration sweep selects the best validation checkpoint and its 800/1600-step masks exceed the 720-update budget. E61b audits an ERM checkpoint. The E44-b accumulation arm clears partial gradients across epoch boundaries and is excluded pending repair. For E62b, the field called `shortcut_rule_ce_risks` is calculated as the negative logarithm of a class-level majority-rule accuracy; this is not predictive cross-entropy and is not used as CE evidence.

The generator's default test parameters are explicitly reported in Sec. 2; unchanged-axis values are 0.8 rather than the training value 0.9. E40/E61 use 512-graph diagnostic subsets, whereas E52/E62 evaluate full 1,500-graph environments. Those metrics must not be pooled. The 240-epoch weight-decay grid uses an absolute budget; its earlier "twice standard" label referred to a different default from the 60-epoch comparison.

Implementation and selection. Each result should be linked to a configuration, dataset hash, split, seed tuple, optimizer state, and metric computation. Record the final partial batch, gradient accumulation, realized optimizer updates, and the release index; do not reconstruct them by multiplying rounded steps per epoch. Hyperparameter and checkpoint selection must identify which validation distribution is used. Test-environment scores must not be used to choose a release time and then reported as an unbiased test estimate. The existing seed counts are not a substitute for a paired analysis.

Reproduction boundary. The accompanying manuscript source reproduces the typeset paper and its figures. Reproducing the experiments additionally requires the training code, raw logs, and configurations. A single aggregate compute-time claim is not made for the growing collection of exper-

iments. The audit and new experiments specified separately are required before treating this draft as a submission-ready empirical validation.

B Supporting Analysis: Statements and Proofs

B.1 Conditional cross-entropy bound

Let $z_\theta(G) \in \mathbb{R}^K$ be the logits on a fixed training set $\{(G_i, y_i)\}_{i=1}^n$, and let $\widehat{L}(\theta) = n^{-1}\sum_i -\log \operatorname{softmax}(z_\theta(G_i))_{y_i}$. For any chosen parameter block B, define $J_{i,B}(t) = \partial z_{\theta_t}(G_i)/\partial \theta_B$ and $m_i(t) = z_{\theta_t}(G_i)_{y_i} - \max_{k \neq y_i} z_{\theta_t}(G_i)_k$. All vector norms below are Euclidean, and matrix norms are operator norms.

Proposition 1 (Conditional CE gradient and displacement bounds)**.** *Suppose that at each update t under consideration the logits are differentiable, $m_i(t) \geq m_t \geq 0$ and $\|J_{i,B}(t)\|_{\mathrm{op}} \leq M_t$ for every training example. Then*

$$\|\nabla_B \widehat{L}(\theta_t)\| \leq A_K M_t e^{-m_t}, \qquad A_K = \sqrt{2}(K-1). \tag{1}$$

For full-batch gradient descent on this CE loss alone, with nonnegative step sizes η_t, this implies

$$\|\theta_B(t_0+H) - \theta_B(t_0)\| \leq A_K \sum_{t=t_0}^{t_0+H-1} \eta_t M_t e^{-m_t}. \tag{2}$$

Other blocks may be updated jointly, provided the stated bounds continue to hold along the resulting trajectory.

The bound applies to any block, including structural and color-input columns. Thus small gradients alone do not identify which features were learned. It bounds motion *after* t_0, not displacement from initialization, and only a summable right-hand side bounds an infinite tail. Exponential margin growth is not assumed for our GNN experiments. Adam, momentum, and additional penalties require their own update analysis. Proofs, partial margin coverage, and a conditional exponential-tail example are in App. B.4.

B.2 An idealized intervention with fixed features and squared loss

For the next two propositions only, let $S \in \mathbb{R}^{d_s}$ and $C \in \mathbb{R}^{d_c}$ be fixed, centered graph-level feature vectors, and $Y \in \mathbb{R}$ a centered scalar response, all with finite second moments. This scalar regression model is separate from the multiclass GNN. The linear predictor is $u^\top S + v^\top C$; its intercept is fixed by centering. The original S and C may be correlated. Suppose $\mu_s I \preceq \Sigma_s = \mathbb{E}[SS^\top] \preceq L_s I$ and $\mu_c I \preceq \Sigma_c = \mathbb{E}[CC^\top] \preceq L_c I$, with $\mu_s, \mu_c > 0$ (redundant coordinates are removed).

Proposition 2 (Independent replacement decouples least squares)**.** *Replace C by an independent copy $\widetilde{C}$ with the same marginal law, independent of (S, Y). Population gradient descent on $L_{\mathrm{ind}}(u, v) = \frac{1}{2}\mathbb{E}[(Y - u^\top S - v^\top \widetilde{C})^2]$ with $0 < \eta \leq 1/\max(L_s, L_c)$ satisfies, for integer $T \geq 0$,*

$$u_T - u_* = (I - \eta\Sigma_s)^T (u_0 - u_*), \qquad v_T = (I - \eta\Sigma_c)^T v_0, \tag{3}$$

where $u_* = \Sigma_s^{-1}\mathbb{E}[SY]$. Writing $q_s = 1 - \eta\mu_s$, $\delta = \|Y - u_*^\top S\|_{L^2}$ and $\epsilon_T = \|Y - u_T^\top S\|_{L^2}$,

$$\|u_T - u_*\| \leq q_s^T \|u_0 - u_*\|, \qquad \epsilon_T \leq \delta + \sqrt{L_s}\, q_s^T \|u_0 - u_*\|. \tag{4}$$

In particular, $v_0 = 0$ implies $v_T = 0$ throughout the idealized warmup.

Here u_* is the best linear structural predictor, not necessarily a perfect or causal predictor: δ includes approximation error and noise. The matrix recursion is not a scalar equality using just the smallest eigenvalue. It gives no CE fitting time and no GNN representation-formation delay.

Proposition 3 (Readout response with a frozen structural predictor)**.** *Under the preceding setup with $v_0 = 0$, restore the original joint law of (S, C, Y) after T warmup steps. Freeze u_T and train only v by population gradient descent on squared loss, with $0 < \eta \leq 1/L_c$. After $H \geq 0$ restoration steps, its value v_H^{post} satisfies*

$$\|v_H^{\mathrm{post}}\| \leq \eta H \sqrt{L_c}\, \epsilon_T, \qquad \|(v_H^{\mathrm{post}})^\top C\|_{L^2} \leq \min\{1, \eta H L_c\}\epsilon_T. \tag{5}$$

The result bounds an additional linear prediction component under the same restored population law, not OOD classification accuracy. It does not describe joint parameter updates in a GNN. Finite-batch node permutation is also not exact independent graph-feature replacement: batch composition and graph size can leave residual dependence (App. B.7). We therefore use warmup as an empirical intervention; neither proposition guarantees its observed accuracy.

B.3 Objects and loss-dependent shortcut risks

The generation parameter $\rho = P(C_{\text{node}} = Y_{\text{class}})$ is a node-level probability. It is not a Pearson correlation, a graph-level Bayes risk, or a fraction of graph-level classification errors. For an arbitrary fixed shortcut statistic $c_i = \psi_c(G_i)$, form equivalence classes $I_a = \{i : c_i = a\}$ on a training set of size n, and let n_{ak} count examples in I_a with class k. The minimum empirical 0–1 risk over *all* deterministic functions of this statistic is exactly

$$\widehat{R}^{\min}_{c,01} = \frac{1}{n} \sum_a \left(|I_a| - \max_k n_{ak} \right). \tag{6}$$

Indeed, a function must predict the same class within each I_a, and choosing that group's majority label minimizes its errors. The fraction of examples belonging to any mixed-label group is generally larger: a group with label counts $(9, 1)$ contains ten conflicting-group members but forces only one classification error. Neither quantity is $1 - \rho$.

For scalar responses y_i with half squared loss, the corresponding unrestricted empirical minimum is

$$\widehat{R}^{\min}_{c,\text{sq}} = \frac{1}{2n} \sum_a \sum_{i \in I_a} (y_i - \bar{y}_a)^2, \qquad \bar{y}_a = |I_a|^{-1} \sum_{i \in I_a} y_i. \tag{7}$$

This follows by minimizing a quadratic separately in each group's prediction. In the population, the corresponding Bayes risks are $\mathbb{E}[1 - \max_k P(Y_{\text{class}} = k \mid C)]$ for classification and $\frac{1}{2}\mathbb{E}[\text{Var}(Y \mid C)]$ for scalar regression. The risk of a restricted linear predictor can exceed these minima. Finite-sample zero error does not establish zero population Bayes risk. Aggregation over multiple nodes can increase graph-level predictiveness without changing ρ; no general identity relates these risks to $1 - \rho$ or $(1 - \rho^2)\,\text{Var}(Y)$.

The CE result below concerns the actual network logits and an arbitrary block of network parameters. The two least-squares propositions instead concern fixed graph-level features and scalar linear coefficients. These are distinct models: there is no assumed equivalence of their losses, features, or optimizers. Centering removes means and fixes the intercept; it does not make correlated features uncorrelated. We assume neither independence nor zero cross-covariance between the original S and C.

B.4 Proof of Proposition 1

Proof. For one example, let $p = \text{softmax}(z)$, with true class y. The logit gradient of CE is $p - e_y$. If its margin is at least $m \geq 0$, then

$$1 - p_y = \frac{\sum_{k \neq y} e^{z_k - z_y}}{1 + \sum_{k \neq y} e^{z_k - z_y}} \leq (K - 1)e^{-m}, \qquad \|p - e_y\|_2 \leq \sqrt{2}(1 - p_y).$$

The second inequality follows from $\sum_{k \neq y} p_k^2 \leq (\sum_{k \neq y} p_k)^2 = (1 - p_y)^2$. By the chain rule, $\nabla_B \ell_i = J_{i,B}^\top (p_i - e_{y_i})$. Taking norms, using the stated Jacobian bound, and averaging over examples gives Eq. 1. The full-batch GD update is $\theta_B(t+1) - \theta_B(t) = -\eta_t \nabla_B \widehat{L}(\theta_t)$. Summing these updates and applying the triangle inequality proves Eq. 2. The calculation needs only the actual Jacobians at update points. For a piecewise differentiable network such as a ReLU GNN, the same inequality holds for the chain-rule Jacobians selected by the implementation when they satisfy the stated norm bound. □

Partial margin coverage. If at time t only a fraction $1 - a_t$ of examples have margin at least $m_t \geq 0$, but every example has $\|J_{i,B}(t)\|_{\text{op}} \leq M_t$, the uniform bound $\|p - e_y\|_2 \leq \sqrt{2}$ on the remaining examples gives

$$\|\nabla_B \widehat{L}(\theta_t)\| \leq \sqrt{2} M_t \big[(1 - a_t)(K - 1)e^{-m_t} + a_t \big]. \tag{8}$$

A nonvanishing uncovered fraction cannot be discarded. Orthogonality of an expected residual would not by itself bound the norm of every sample gradient or its variance.

A conditional exponential-tail example. If, in addition, for all integers $k \geq 0$, $m_{t_0+k} \geq m_0 + ck$, $M_{t_0+k} \leq M_0 e^{ak}$, $\eta_{t_0+k} \leq \bar{\eta}$, and $c > a \geq 0$, then

$$\|\theta_B(t_0 + H) - \theta_B(t_0)\| \leq A_K \bar{\eta} M_0 e^{-m_0} \frac{1 - e^{-(c-a)H}}{1 - e^{-(c-a)}}. \tag{9}$$

The displacement from the fixed starting time t_0 need not vanish as H grows; it has a finite upper bound. Starting instead at $t_0 + q$ multiplies the remaining infinite-tail bound by $e^{-(c-a)q}$. This corrects the distinction between a small *late tail* and a vanishing total displacement. Linear-in-time margin growth is an additional premise, not an implication of CE interpolation. For example, the logarithmic margin growth of separable linear logistic regression (Soudry et al., 2018) does not yield this exponential-tail conclusion. We have not verified the margin and Jacobian envelopes for the GNN runs.

What the bound does not identify. It is possible to have small CE gradients after learning structure, after learning a shortcut, or after learning both. A parameter block identified by first-layer feature columns is not the entirety of a structural representation: adjacency and shared layers also contribute. The calculation neither asserts curvature collapse nor proves that memorization bypasses the structural pathway. Small curvature in one parameter block would not by itself bound a different block's gradient. Reparameterizing a predictor can change gradient norms without changing its predictions, so a gradient-norm ratio is not an invariant allocation measure. SGD requires control of its sampled updates; Adam and momentum depend on additional state. The optimizer forks concern those empirical dynamics, not an extension of Proposition 1.

B.5 Proof of Proposition 2

Proof. The independent replacement has mean zero and is independent of (S, Y). Consequently $\mathbb{E}[S\widetilde{C}^\top] = 0$ and $\mathbb{E}[\widetilde{C}Y] = 0$. Expanding the population half squared loss gives the exact decomposition

$$L_{\text{ind}}(u, v) = \tfrac{1}{2}\mathbb{E}[(Y - u^\top S)^2] + \tfrac{1}{2} v^\top \Sigma_c v.$$

Let $b_s = \mathbb{E}[SY]$. Its gradients are $\nabla_u L_{\text{ind}} = \Sigma_s u - b_s$ and $\nabla_v L_{\text{ind}} = \Sigma_c v$. Since $u_* = \Sigma_s^{-1} b_s$, the GD recursions are

$$u_{t+1} - u_* = (I - \eta\Sigma_s)(u_t - u_*), \qquad v_{t+1} = (I - \eta\Sigma_c) v_t.$$

Iteration proves Eq. 3. Because $0 < \eta \leq 1/\max(L_s, L_c)$, all eigenvalues of these iteration matrices lie in $[0, 1)$, and their operator norms are at most $1 - \eta\mu_s$ and $1 - \eta\mu_c$, respectively. Hence $\|u_T - u_*\| \leq q_s^T \|u_0 - u_*\|$ and $\|v_T\| \leq (1 - \eta\mu_c)^T \|v_0\|$. Finally the triangle inequality in L^2 gives

$$\epsilon_T \leq \|Y - u_*^\top S\|_{L^2} + \|(u_T - u_*)^\top S\|_{L^2} \leq \delta + \sqrt{L_s}\|u_T - u_*\|.$$

Substitution proves Eq. 4. Zero v_0 remains zero under every population update. □

A sufficient duration within this model. Set $D = \|u_0 - u_*\|$. For a desired coefficient error $\varepsilon > 0$ and $D > \varepsilon$, it suffices to take $T \geq \lceil (\eta\mu_s)^{-1} \log(D/\varepsilon) \rceil$, since $(1 - \eta\mu_s)^T \leq e^{-\eta\mu_s T}$. If $D \leq \varepsilon$, $T = 0$ already suffices. This controls coefficient error, not multiclass accuracy or the formation of a nonlinear motif representation. Observed gradient norms cannot be substituted for μ_s without a separate model connecting those quantities.

Singular feature covariances and finite batches. Centered one-hot or histogram features often have redundant coordinates. The stated positive-definiteness assumption can be met by restricting each fixed feature vector to its covariance range. Components in a null space are unidentifiable and do not contribute to predictions almost surely. The displayed coefficient contraction applies on the restricted space. These are gradients of an expected quadratic objective. With finite-batch stochastic gradients, $v_0 = 0$ need not remain zero in a realized run, even under genuinely independent replacement; additional variance and concentration analysis would be needed.

B.6 Proof of Proposition 3

Proof. Let $r_T = Y - u_T^\top S$ and $h_T = \mathbb{E}[Cr_T]$ under the original joint population law. With u_T frozen, the restored objective is $\frac{1}{2}\mathbb{E}[(r_T - v^\top C)^2]$, whose gradient is $\Sigma_c v - h_T$. Starting from $v_T = 0$, write $A = I - \eta\Sigma_c$. Then

$$v_H^{\text{post}} = \eta \sum_{j=0}^{H-1} A^j h_T.$$

The step-size condition implies $\|A\|_{\text{op}} \le 1$. For every unit vector b, Cauchy–Schwarz gives

$$|b^\top h_T| \le \sqrt{\mathbb{E}[(b^\top C)^2]}\sqrt{\mathbb{E}[r_T^2]} \le \sqrt{L_c}\epsilon_T.$$

Taking the supremum over b and summing the recursion proves the coefficient bound in Eq. 5. The prediction bound follows from $\|v^\top C\|_{L^2}^2 = v^\top \Sigma_c v \le L_c\|v\|^2$. For the uniform bound, set $w = \Sigma_c^{-1/2} h_T$. Expanding the squared residual at its optimal linear coefficient gives

$$0 \le \mathbb{E}[(r_T - h_T^\top \Sigma_c^{-1} C)^2] = \epsilon_T^2 - h_T^\top \Sigma_c^{-1} h_T.$$

Thus $\|w\| \le \epsilon_T$. The exact recursion and the fact that A commutes with Σ_c give

$$\Sigma_c^{1/2} v_H^{\text{post}} = (I - A^H)w.$$

Every eigenvalue of A lies in $[0, 1)$, so $\|I - A^H\|_{\text{op}} \le 1$ and $\|(v_H^{\text{post}})^\top C\|_{L^2} \le \epsilon_T$. Combining this with the finite-step bound proves the minimum in Eq. 5, including $H = 0$. No original-channel orthogonality is used. □

Finite-horizon interpretation. For this restricted model, a sufficient condition for the restored component to have L^2 norm at most $b > 0$ is $\eta H L_c \epsilon_T \le b$. The sharper uniform bound shows that the restored component remains bounded by the structural residual even at long horizons. It is not a classification margin guarantee. If $\epsilon_T = 0$, $h_T = 0$ and the restored component stays zero. If $\epsilon_T > 0$, the restored channel may learn residual signal. Jointly updating the structural representation, using CE or Adam, or changing the test distribution falls outside this proposition.

B.7 Finite-batch permutation is not independent replacement

Fix a batch with N nodes, color vectors $x_1^C, \ldots, x_N^C$, and graph node sets V_g of size n_g. Under a uniform permutation π of all batch nodes,

$$\mathbb{E}_\pi\left[\sum_{j \in V_g} x_{\pi(j)}^C \,\middle|\, B\right] = n_g \bar{x}_B^C, \qquad \bar{x}_B^C = \frac{1}{N}\sum_{j=1}^N x_j^C. \tag{10}$$

Each destination receives each source with probability $1/N$, proving the identity. A mean-pooled histogram has conditional mean $\bar{x}_B^C$; a sum-pooled histogram additionally retains graph-size dependence. The source pool contains the target graph's own nodes. For independent graphs of fixed sizes, writing $p_y = \mathbb{E}[x_j^C \mid Y_g = y]$ for a node of the target graph and $\bar{p}$ for a node of any other graph gives

$$\mathbb{E}[\bar{x}_B^C \mid Y_g = y] = \frac{n_g}{N} p_y + \left(1 - \frac{n_g}{N}\right)\bar{p}.$$

Thus even the conditional mean need not be label-independent. Batch stratification and label-dependent graph sizes can add further dependence. Matching conditional means would still not imply independence of full feature vectors or nonlinear GNN embeddings.

For a general corrupted feature C' with mean zero, define $\Gamma = \mathbb{E}[S(C')^\top]$, $b_c' = \mathbb{E}[C'Y]$, and $\Sigma_c' = \mathbb{E}[C'(C')^\top]$. The squared-loss gradients are

$$\nabla_u L = \Sigma_s u - \mathbb{E}[SY] + \Gamma v, \qquad \nabla_v L = \Sigma_c' v + \Gamma^\top u - b_c'.$$

These extra terms disappear in Proposition 2, but not automatically for the implemented permutation. E59 measures residual color predictiveness in the actual warmup batches. Its near-chance probe supports effective empirical suppression; it does not establish the exact population independence premise or a universal $O(N^{-1/2})$ error bound.

C Diagnostic and Auxiliary Observations

Color rule and accuracy sweep. At fixed $b_{\text{struct}} = 0.9$, reported worst-environment accuracy is 98.4%, 93.1%, 15.4%, and 8.8% for $\rho = 0.3, 0.5, 0.7, 0.9$. A two-seed left-arm sweep at $\rho = 0.15, 0.2, 0.25$ gives 14.0%, 59.2%, and 97.9%. These are empirical response bands, not estimated phase-transition thresholds. At $\rho = 0.1$, the per-node mutual information under balanced labels is approximately 0.150 nats, versus 0.0025 at 0.3. For graphs containing multiple nodes, those values do not directly give the predictive information in the full color histogram.

Gradient traces. First-layer color-to-degree gradient-norm ratios are approximately 1.5–1.7 after ordinary ERM fitting. During permutation, the ratio on the training inputs is approximately 0.8–1.0; an unpermuted counterfactual probe gives 1.4–2.5. These are norms of different input-column blocks, under a fixed feature scaling. They neither measure all structural computations nor certify the margin–Jacobian assumptions of App. B. Small gradients after fitting can follow learning either channel.

Representation probes. For the $\rho = 0.1$ experiment, a fresh linear classifier is trained on a frozen encoder using 2,000 training graphs balanced by majority color within each class, with 500 held-out training graphs for validation. The reported shifted-environment accuracies are 0.03–0.14 for the ERM encoder and 0.61–0.96 for the suppressed-prefix encoder. Balancing majority color does not equalize full color-histogram distributions. These results compare recoverability under this protocol; they cannot establish absence of structural information. A fresh random encoder also gives low shifted-environment accuracy under the reported head-training protocol.

Input ablation and architecture. Zeroing colors at test time in a single-head model can restore structure-shift accuracy to approximately 97–98% from 0–7%. In a separately trained two-head model at $\rho = 0.1$, accuracy changes from 2.1% to 2.4% under input zeroing. This is an architecture comparison, not a localization of the original model's failure: zeroing colors changes both message passing and readout inputs, and retraining changes the representation.

P2 norm-matched forks: paired intervals. Seed-paired differences in final anti-environment accuracy over ten data seeds (bootstrap 95% intervals, 10,000 resamples): F0$-$F1 $= +0.716$ $[0.669, 0.760]$; F0$-$F5 (shock-only vs. reference) $= +0.644$ $[0.593, 0.692]$; F3$-$F1 (global norm match vs. v-reset) $= +0.025$ $[-0.040, 0.087]$; F2$-$F1 (full reset vs. v-reset) $= +0.448$ $[0.383, 0.528]$; F0$-$F2 $= +0.268$ $[0.186, 0.344]$. Matching rescales the applied parameter update $\theta \leftarrow \theta + \alpha d$ with $\alpha = \|\Delta\|_{\text{target}}/\|d\|$ after the normal Adam step; the scale factor is not folded into the moment estimates, and zero-norm steps that cannot match a positive target are logged and reported, not silently clipped. All branches read the same pre-generated data stream as their F0/F1 references (`p2_forks.jsonl`, `p2_summary.json`).

Post-restoration trajectory diagnostic. A seed-0 run with $T_w = 200$ and 1,240 continuation updates ends at 0.836 anti accuracy, compared with 0.998 for permanent permutation. Its probe cross-entropy also varies over training. The quantity $\log K - \text{CE}_{\text{probe}}$ is a score relative to uniform prediction; it is not mutual information without additional assumptions. Its rolling variation does not by itself establish feature transfer or optimizer-state damage. We use the recorded endpoint as supplementary horizon evidence and do not infer a mechanism from the diagnostic.

Additional scope limits. The initial Cora analysis correlates integer class indices with degree and principal components. Those statistics change under relabeling and do not validate the main intervention. They are not treated as real-data support here. The weight-decay grid varies shortcut strength and penalty at a fixed 240-epoch horizon; it does not measure a norm hierarchy or derive a recovery time. These auxiliary studies require their own controls before becoming central claims.

External task, including a negative result (P4). GOOD concept-shift split (verbatim protocol, fixed seed; inference scope: optimization seeds only); suppression re-implemented as per-graph palette resampling with the exact inverse $\text{intensity} = \max_{\text{RGB}}$ (unit-tested; intensity, topology, labels preserved); four arms to T=6,800, batch 256, P1 stream/fork protocol; ten optimization seeds. Final OOD accuracy: ERM 0.868 ± 0.003, release$_{200}$ 0.868 ± 0.003, release$_{800}$ 0.869 ± 0.003,

continuous re-colorization 0.258 ± 0.021. Seed-paired differences: ERM$-$suppression $= +0.610$ [0.598, 0.624]; ERM$-$release$_{200} = +0.000$ [-0.002, 0.002]; ERM$-$release$_{800} = -0.001$ [-0.003, 0.001]. The opposite ordering from RedHouse is attributable to the split's diverse shortcut proportions (0.75–0.95; the mixture regime), plus two obstacles to suppression here: intensity-only inputs leave training CE at chance under this architecture, and palette resampling injects input-norm variance ($\|\text{palette}\|_2 \in [1.0, 1.5]$). The historical 20-epoch pilot's ERM collapse (0.1417 at $\sim$2,780 updates) is an undertrained artifact not reproduced at the full horizon; the small `good_color.json` hybrid pilot (2 seeds; match probability $b+(1-b)/K$) is not pooled (`p4_runs.jsonl`, `p4_summary.json`, `p4_ceiling.json`).

D HISTORICAL LOW-LABEL COMPARISONS

The 10%-label experiments compare methods under a short schedule on a homogeneous shortcut distribution. They answer an accuracy question about this setting, not the restoration question of the main text. The color-aware permutation arms use information unavailable to a generic environment-based objective. "Oracle environment" means knowledge of the background-family tag, not knowledge of the color columns. These privileges are distinct. GroupDRO uses class–environment groups

Table 6: REDHOUSE, 10% labels, mean $\pm$ std of worst-environment accuracy over 3 seeds (batch 256; total budget $\approx$120 optimizer steps, so the T_w=200 permutation arm is mask-throughout; Sec. 3). IRM/V-REx/GroupDRO use oracle environments partitioned by background topology family (BA/ER/WS); CIGA uses the reported reimplementation (App. D). The rows differ in privileged information and should be interpreted under this particular training distribution.

Method	struct	feat	both	anti	worst
ERM	98.7±0.9	41.1±1.1	33.1±0.1	14.5±3.6	14.5±3.6
IRM (env oracle, λ=100, anneal 0)	66.0±6.9	40.0±3.3	34.0±1.3	22.5±3.5	22.5±3.5
V-REx (env oracle, β=100)	98.7±0.3	41.3±2.1	34.8±0.6	12.9±4.7	12.9±4.7
GroupDRO (env oracle)	99.3±0.2	35.9±3.5	33.9±1.3	13.2±0.6	13.2±0.6
CIGA (reimpl.)	99.8±0.1	37.4±1.2	34.7±1.3	7.2±1.2	7.2±1.2
{IRM/V-REx} (15/6-config sweeps)	worst 8.1–12.5 across all				$\leq$12.5
DIR	99.8±0.2	38.3±1.3	34.9±1.6	7.5±1.1	7.5±1.1
GSAT	99.6±0.2	39.5±0.9	33.5±0.2	13.2±3.9	13.2±3.9
UniGOOD (reimpl.)	99.7±0.0	39.0±2.8	34.1±0.7	9.4±1.3	9.4±1.3
permutation throughout (ours)†	92.9±6.4	96.8±2.5	92.4±6.4	89.7±7.6	89.7±7.6
permutation + consistency (ours)†	**95.1±0.4**	**98.0±0.2**	**95.0±0.3**	**92.9±0.7**	**92.9±0.7**

†Permutation arms are mask-*throughout* runs (120-step budget never releases; released-warmup: Sec. 3; protocol inventory: App. A). IRM/V-REx: penalty active from step 0 (λ=β=100; E56); the sweep row subsumes other weights/anneals.

with exponentiated group weights. The reported CIGA implementation uses attention-weighted subgraph pooling, class-conditional contrastive loss, and an attention regularizer on the GIN backbone. Exact compatibility with each reference implementation requires a code audit. The original IRM annealing schedule exceeded the short training budget; the displayed rerun activates the penalty at step zero. This correction is necessary but does not establish exhaustive hyperparameter optimization.

Full-label 60/240-epoch controls use a different protocol. In those controls, homogeneous topology-grouped runs remain weak, while a diverse mixture supports strong ERM, CIGA, and DIR results. Consequently the historical table should not be read as a general ranking of robust learning methods or evidence that input suppression is necessary.