KNOWN-SHORTCUT INTERVENTIONS UNDER MIXED GRAPH DISTRIBUTION SHIFTS: A DIAGNOSTIC STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph classifiers can exploit spurious correlations in both background topology and node attributes. We study how knowledge of a shift-prone feature block affects robustness under such distribution shifts. We introduce REDHOUSE, a synthetic graph-classification benchmark with independently controllable correlations between the label and the background graph family or node colors. Under a protocol with 300 labeled training graphs, targeted color-block randomization substantially improves worst-environment accuracy over the evaluated baselines. A prior-matched comparison shows that a simple GIN with the same finite-warmup intervention (T_w=100) achieves $95.7\pm2.6\%$ worst-environment accuracy across ten seeds, compared with $93.6\pm4.2\%$ for the full MIXEDSTABLE framework; these results do not establish an advantage of the full framework. We distinguish persistent randomization from finite warmup followed by shortcut release: the main low-label setting primarily evaluates sustained shortcut suppression, whereas longer post-release training reveals substantial performance decay and seed-dependent failures. Additional experiments on external benchmark variants show heterogeneous outcomes, including failures of the full framework. Our findings identify targeted intervention on a known nuisance block as the main source of the observed REDHOUSE gains and motivate evaluations that match prior information, training budgets, and checkpoint selection. They do not establish that a short warmup reliably yields lasting invariance or that mixed shifts necessarily cause stronger shortcut dependence than feature shifts alone.

1 INTRODUCTION

Graph classifiers can exploit spurious correlations along two distinct channels: the background topology on which a functional motif appears (base-graph family), and the node-attribute distribution that annotates it (node colors). Both correlations are common in training data; at test time either or both marginals may change. These are three separate statements—a training-side shortcut may exist, a test-side factor may shift, and the two channels may interact—and this paper does not assume any of them is established in general. We therefore study them in a setting where each factor is separately controllable.

Comparisons between robustness methods can confound the effect of a learning objective with the information supplied by its augmentations. In particular, an intervention that targets a known nuisance block gives the learner information unavailable to an unmodified baseline. A diagnostic evaluation should therefore compare methods under matched prior information and distinguish persistent shortcut suppression from transient interventions during training.

We ask three questions. First, how much of the observed robustness gain is explained by targeted intervention on a known nuisance block? Second, does a finite intervention remain effective after the original shortcut is restored? Third, which additional components, if any, improve performance under the same prior and evaluation protocol?

Contributions. *(1)* We introduce REDHOUSE, a controlled diagnostic benchmark with independently adjustable background-family and color-label correlations (Sec. 3). *(2)* We evaluate targeted shortcut interventions using prior-matched simple baselines and distinguish persistent randomiza-

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.

tion from finite warmup followed by release (Secs. 4.1–5). *(3)* We analyze the incremental value and failure modes of a composite robustness framework (MIXEDSTABLE) through component ablations and external benchmark variants (Sec. 5, Apps. D.9–D.8). The headline findings: a simple GIN with the same finite-warmup intervention (T_w=100) shows no established disadvantage against the full framework on REDHOUSE (95.7±2.6 vs. 93.6±4.2 worst-environment across ten seeds, Welch p=0.20); external variants show heterogeneous outcomes including outright failures of the full framework; and high accuracy during the intervention does not imply reliable preservation after the shortcut returns.

2 RELATED WORK

Graph OOD benchmarks and methods. GOOD (Gui et al., 2022) builds separate covariate-shift splits (structure or feature) and concept-shift splits; Spurious-Motif (Wu et al., 2022b) controls a single structure-level bias. Supervised invariant-subgraph methods (Wu et al., 2022b; Miao et al., 2022; Chen et al., 2022) and causal interventions (Wu et al., 2022a; 2024; Fan et al., 2023) address structure-level shifts; recent extensions cover node-level causality (Lu et al., 2026) and supervised HSIC penalization (Zhang et al., 2026). Environment-based invariance (Arjovsky et al., 2019; Krueger et al., 2021) transfers to graphs only with environment annotations. OOD-GCL (Li et al., 2024) disentangles latent channels for a single shift source; UniGOOD (Liao et al., 2026) unifies OOD generalization and detection with label-free invariant-subgraph generation. None of these evaluations isolates the two covariate-shift channels with independently adjustable correlations, which is the axis REDHOUSE provides; our diagnostic question—how much of a method's measured gain is the information in its targeted augmentation—applies to all of them.

Stable learning. Stable learning removes spurious correlations via variable balancing (Cui & Athey, 2022; Shen et al., 2018; 2020), deepened end-to-end (Zhang et al., 2021) with covariate-shift theory (Xu et al., 2021; 2022); S-GNN brings RFF-space decorrelation to supervised node classification, and VIVACE (Li et al., 2025) corrects variant-subgraph spuriousness by reweighting. Our diagnostic re-evaluates this family's assumptions in a setting where a covariate-based selection criterion can be shown to fail in population (App. A.3), and our experiments attribute the measured gains to the augmentation prior rather than the selection machinery.

Environment-free invariance and pretrained encoders. Environment inference (Creager et al., 2021; Lin et al., 2022) and norm-guided IRM-Free (Qiu et al., 2025) reconstruct or replace environments; the dual-channel consistency in MIXEDSTABLE derives invariance from targeted augmentation families instead. A GFM-OOD survey (Anonymous, 2026a) frames graph foundation model robustness as an open evaluation problem; our channel-sensitivity analysis of pretrained GIN encoders (App. C) is a step toward per-channel evaluation. SAM (Foret et al., 2021) and its graph variants (Luo et al., 2024; Qiu et al., 2026) seek flat minima; the feature+weight-perturbed SSL variant used in our probe experiments is described in the companion paper. *Nomenclature note:* an unrelated method also called "GraphSAM" (Wang et al., 2024) accelerates SAM for molecular graph transformers via gradient reuse; any same-named objective in this paper's experiments is a distinct SAM-style SSL method.

3 CONTROLLED BENCHMARK AND PROTOCOLS

Existing graph OOD evaluations vary one distribution dimension at a time: GOOD (Gui et al., 2022) builds separate covariate-shift splits (structure or feature) and concept-shift splits; Spurious-Motif (Wu et al., 2022b) controls a single structure-level bias. Evaluating targeted interventions under mixed shifts requires a benchmark whose two correlation parameters can be adjusted *independently* (we do not claim the two observed channels are statistically independent).

REDHOUSE: generation. Each graph's label is determined solely by its *motif* class (K=3: house / cycle / crane; 5–6 nodes). A base graph of 10–20 nodes is drawn from a family ($B \in \{$BA, ER, WS$\}$) and the motif is attached by a single bridge edge to a uniformly random base node. Classes

are balanced ($y = i \bmod K$). Given the label y:

$$\mathbb{P}(B = b \mid Y = y) = \begin{cases} b_{\text{struct}}, & b = \text{family}(y), \\ (1 - b_{\text{struct}})/(K-1), & \text{otherwise} \end{cases}$$

(the $(K-1)$ non-matching families are sampled uniformly), and for each node v, independently,

$$\mathbb{P}(C_v = c \mid Y = y) = \begin{cases} b_{\text{feat}}, & c = y, \\ (1 - b_{\text{feat}})/(K-1), & c \neq y. \end{cases}$$

Base family and node colors use independent random draws conditioned on Y (they are not coupled to each other beyond their shared dependence on the label). Training uses $b_{\text{struct}}{=}b_{\text{feat}}{=}0.9$. Node features concatenate a degree one-hot (truncated at $\min(d, 6)$, 7 dims) and a color one-hot (3 dims). Train / validation / test splits are generated from a single seeded generator stream with disjoint graph indices; node indices follow the construction order (base nodes first, then motif nodes) and are not reshuffled; the ground-truth motif mask is exported for evaluation only and is never an input. Four test environments vary the two correlations independently:

$$(b_{\text{struct}}, b_{\text{feat}}) \in \{(1/3,\ 0.9),\ (0.9,\ 1/3),\ (1/3,\ 1/3),\ (0.2,\ 0.2)\},$$

i.e., *struct-shift*, *feat-shift*, *both-shift*, and *anti* (anti-correlated spurious).

Metrics. Accuracy per environment; worst-environment accuracy (robustness; computed per seed first, then aggregated across seeds); label-efficiency curves over $\rho \in \{2\%, 5\%, 10\%, 30\%, 100\%\}$. Dataset statistics: 3,000 training, 300 validation, and 1,500 test graphs per environment; 10–27 nodes per graph; feature dimension 10.

3.1 Targeted feature intervention

The intervention evaluated throughout this paper permutes the *color-block feature rows across nodes*. Formally, within each training minibatch, let N be the total number of nodes and let $x_{v,7:10}$ denote the color one-hot row of node v. A permutation π is drawn uniformly over the N nodes of the minibatch, and the color rows are reassigned: $x'_{v,7:10} \leftarrow x_{\pi(v),7:10}$ for all v (fraction p of nodes; $p{=}1$ throughout). Concretely: (1) the permutation acts on the *node axis* of the color block only; (2) it is *cross-graph*: π ranges over all nodes in the minibatch, so color rows are exchanged both within and across graphs; (3) each node participates once (whole-batch permutation, not per-graph independent draws); (4) π is *resampled every training step*; (5) the degree one-hot block (dims 0:7) and the graph structure are untouched; (6) *evaluation never applies the intervention*—test inputs are unmodified. The effect is that each node's color becomes an exchangeable draw from the minibatch's empirical color distribution, severing the color–label association while leaving each node's degree and the motif's connectivity intact.

Prior boundary. Identifying a contiguous one-hot block specifies its input location; it does not establish that the block is non-causal or shift-prone. Our targeted intervention assumes this additional knowledge.

Regimes. **Persistent color randomization**: the intervention is applied at every training step. **Finite warmup and release**: the intervention runs for the first T_w steps, after which the original view returns (*release of the original view*). Under the low-label main protocol ($\rho{=}10\%$, $\sim$120 optimizer steps), the composite framework's default $T_w{=}800$ applies the intervention throughout training (persistent regime); a finite-warmup variant with $T_w{=}100$ occupies $\sim$83% of the training updates in this protocol. Under the 100%-label schedule (720 steps), $T_w{=}100$ occupies $\sim$14%. **Block-agnostic intervention**: permuting all feature blocks without knowing which one is shift-prone (Table 2).

Intervention validity. A color-histogram probe (ridge one-vs-all, 100 resplits of one pooled-node permutation draw) predicts the label at 100.0% from mean color one-hots before intervention and $33.8{\pm}4.1\%$ (chance $= 33.3\%$) after permutation; the degree-histogram probe retains $83.5{\pm}3.2\%$ and a motif-restricted degree probe 100%—the intervention substantially reduces label predictability from the evaluated color-histogram probe while the structure probe survives (details and caveats in App. D.4).

4 INTERVENTIONS AND REFERENCE METHODS

This section describes the interventions and reference methods evaluated in Sec. 5. Formal statements that motivated earlier versions of the method stack—a risk-invariance lemma, an analysis of pairwise HSIC monitoring, and an impossibility counterexample for covariance-based stable-set selection—are relegated to App. A as background facts; the main text makes no recovery or invariance claim that depends on them.

4.1 TARGETED AND REFERENCE INTERVENTIONS

All methods share the GIN backbone of Sec. 3 and are evaluated under identical budgets and checkpoint rules. The *targeted intervention* permutes the color-block feature rows across nodes (formal definition in Sec. 3.1); it is used in two regimes:

Persistent color randomization. The intervention is applied at every training step. This is the operative regime of the low-label main protocol: with ρ=10% labels the training budget is $\sim$120 optimizer steps, so the default schedule (T_w=800) applies the intervention throughout.

Finite warmup and release. The intervention runs for the first T_w steps only, after which the original view returns (*release of the original view*). Under the 100%-label schedule (720 steps, T_w=100) the intervention occupies the first $\sim$14% of updates; under the low-label protocol a finite-warmup variant with T_w=100 occupies $\sim$83% of the $\sim$120 total updates. The two regimes are reported separately throughout.

MIXEDSTABLE (reference composite framework). The composite framework under diagnosis combines: (a) stable-subgraph attention with DIR-style interventional consistency; (b) a hard-concrete dimension mask with sample reweighting (capacity-control heuristics; App. A); (c) dual-channel pseudo-environment consistency with prediction-level symmetric KL, representation alignment, and a risk-variance penalty; and (d) persistent color randomization of the original view during warmup—under the main protocol, throughout training. *Prior knowledge used.* The structure augmentation is uniform edge dropout using no ground-truth annotation. The feature augmentation and the warmup both assume knowledge of *which feature block is shift-prone*; block-agnostic and dropout-based controls that do not use this knowledge are compared in Table 2. Identifying a contiguous one-hot block specifies its input location; it does not establish that the block is non-causal or shift-prone. Our targeted intervention assumes this additional knowledge.

Scope of the invariance objective. Our augmentation objective does not provide an unconditional guarantee of mixed-shift invariance. Such a guarantee would require label-preserving transformations, adequate coverage of the target distribution, a realizable predictor, and regularizers compatible with that predictor. These properties are not established for the implemented edge-dropout and feature-randomization operators. We therefore evaluate the resulting robustness empirically; App. A records the assumptions under which a formal statement would hold, for reference.

Whether the consistency regularization *preserves* the structural solution once the shortcut returns is evaluated with a same-checkpoint fork design (common warmup, then release $\pm$consistency vs. persistent randomization; App. D.3).

The full objective adds supervised CE on the (weighted, few-label) batch, $\mathcal{L} = \mathrm{CE}_w + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}} + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}} + \lambda_{\ell_0}\|M\|_0 + \lambda_a\|a\|_1$, trained with block-coordinate alternation for (θ, M, a) and w (step outline in Appendix A.5).

5 DIAGNOSTIC EXPERIMENTS

5.1 SETUP

Benchmarks. REDHOUSE (Sec. 3), plus external benchmark variants: GOOD-Motif, GOOD-CMNIST (v1/v3), and a GOOD+Color hybrid (Sec. 5.5).

Methods. Supervised: ERM (GIN), IRM (Arjovsky et al., 2019), V-REx (Krueger et al., 2021) (both with oracle environment labels), DIR (Wu et al., 2022b), GSAT (Miao et al., 2022). Label-free invariant: UniGOOD (reimpl.) (Liao et al., 2026). Ours: HSIC-SSL, SPARSESTABLE, MIXED-

Table 1: REDHOUSE mixed shift, $\rho = 10\%$ labels (300 labeled graphs), 60 epochs, mean $\pm$ std over 3 seeds (per-environment values in App. D). The evaluated methods perform poorly under this protocol on the feature-shifted environments; targeted intervention (persistent randomization, in MIXEDSTABLE or applied to plain GIN) holds all four—attribution in Table 2.

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)	98.7 ± 0.9	41.1 ± 1.1	33.1 ± 0.1	14.5 ± 3.6	14.5 ± 3.6
V-REx (env oracle)	97.6 ± 1.6	38.6 ± 1.0	34.3 ± 1.3	12.4 ± 5.6	12.4 ± 5.6
DIR	99.7 ± 0.2	38.4 ± 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	39.0	34.1	9.4	9.4
HSIC-SSL (ours, 2a)	54.2	44.6	32.3	24.6	24.6
SPARSESTABLE (ours, 2b)	97.8	36.7	35.4	11.8	11.8
MIXEDSTABLE + persistent color rand.	95.1 ± 0.4	**98.0** ± 0.2	95.0 ± 0.3	**92.9** ± 0.7	**92.9** ± 0.7
GIN + finite warmup (T_w=100)	97.2 ± 1.8	98.9 ± 0.3	97.5 ± 1.8	96.1 ± 2.2	96.1 ± 2.2
GIN + persistent color rand.	96.3 ± 1.5	98.8 ± 0.4	97.0 ± 1.6	95.5 ± 2.0	95.5 ± 2.0

STABLE (reference composite framework, Sec. 4.1). All share the same GIN backbone (3 layers, hidden 64), Adam, tuned lr $\in [10^{-4}, 10^{-3}]$.

Protocols. Low-label: $\rho = 10\%$ labeled graphs (300), 60 epochs ($\sim$120 optimizer steps), final-epoch and IID-validation-selected read-outs reported separately where they differ. Full-label: 3,000 graphs, 720 steps.

5.2 Q1: HOW MUCH OF THE GAIN IS THE TARGETED INTERVENTION?

Table 1: the evaluated methods perform poorly under this protocol—IRM/V-REx with oracle environments, DIR/GSAT subgraph selection, and the UniGOOD-style spectral learner all drop to 7–41% once the feature channel shifts. Training without targeted intervention yields low OOD accuracy despite a high-performing color-removed control (90.8 ± 1.2 in Table 2; an earlier full-data oracle run with colors deleted reported 99.7 under a different checkpoint rule and 100%-label protocol, so the two numbers are not comparable). The decisive controls are the bottom rows: plain GIN with finite warmup (T_w=100 of $\sim$120 updates, $\sim$83%) reaches 96.1 ± 2.2 and with persistent randomization 95.5 ± 2.0. Targeted intervention on the known nuisance block is the main source of the observed gains (full attribution below).

Prior-matched attribution (Table 2): plain GIN+linear-head baselines under the exact Table 1 protocol with the same augmentation budget but different priors. The evaluated block-agnostic variants exhibit substantially lower accuracy: permuting *all* one-hot blocks stays at 19.1 ± 6.8; random feature dropout at 17.9 ± 1.1; zeroing the color block (an oracle) reaches 90.8 ± 1.2. Persistent color randomization works as well as the finite-warmup variant (95.5 ± 2.0 vs. 96.1 ± 2.2), and V-REx with oracle environments adds nothing on top of the intervention (84.6 ± 11.0). A 10-seed statistical study (App. D.7) extends the headline comparison to finite-warmup GIN (T_w=100): 95.7 ± 2.6 vs. full MIXEDSTABLE 93.6 ± 4.2 (mean diff $+2.1$, Welch 95% CI $[-1.2, +5.4]$; Welch p=0.20; Wilcoxon p=0.32)—these results do not establish an advantage of the full framework, and the point estimate favors the simple control. An earlier 3-seed run of IRM on top of the intervention reported 96.2 ± 2.6 because its IRM penalty was implemented with a step-50 activation buffer (in effect mostly off); with the penalty active from step 0, the 10-seed estimate is 81.2 ± 7.5 ($p < 10^{-3}$ vs. the intervention alone)—we report the always-on version as IRM's behavior and flag the implementation difference explicitly.

5.3 Q2: DOES A FINITE INTERVENTION SURVIVE RELEASE?

Same-checkpoint fork. To isolate what happens after release from warmup-quality confounds, all arms share an identical warmup trajectory (100% labels: common warmup T_w=100 steps, then fork from the same checkpoint—model and optimizer state—into: original view returns without consistency; original view returns with the consistency regularizer; persistent randomization continues;

Table 2: Prior-matched controls: plain GIN+linear head, exact Table 1 protocol (ρ=10% labels, 300 graphs, 60 epochs, 3 seeds unless noted, mean ± std, best-IID-val read-out for this table). The finite-warmup variant uses T_w=100 of ∼120 optimizer steps (∼83% of the training updates). The block-specific prior alone suffices; the evaluated block-agnostic variants exhibit substantially lower accuracy. The IRM row uses the always-on penalty at n=10 (App. D.7); its step-buffered 3-seed variant reached 96.2±2.6 but with the penalty effectively inactive.

Variant	struct	feat	both	anti	worst
ERM (reference)	98.3 ± 0.8	34.9 ± 1.8	35.6 ± 3.2	14.2 ± 2.1	14.2 ± 2.1
GIN + finite warmup (T_w=100)	97.2 ± 1.8	98.9 ± 0.3	97.5 ± 1.8	96.1 ± 2.2	96.1 ± 2.2
GIN + block-agnostic perm	98.6 ± 1.2	47.9 ± 4.9	40.7 ± 5.3	19.1 ± 6.8	19.1 ± 6.8
GIN + feat dropout 0.3	96.6 ± 2.9	37.7 ± 3.1	35.6 ± 1.1	17.9 ± 1.1	17.9 ± 1.1
GIN + color zeroed (oracle)	95.5 ± 4.0	98.0 ± 0.4	95.6 ± 1.0	92.6 ± 3.4	90.8 ± 1.2
GIN + persistent color rand.	96.3 ± 1.5	98.8 ± 0.4	97.0 ± 1.6	95.5 ± 2.0	95.5 ± 2.0
IRM (oracle, penalty from step 0; n=10)	–	–	–	81.2 ± 7.5	81.2 ± 7.5
V-REx (env oracle) + intervention	96.1 ± 2.2	94.6 ± 5.7	89.6 ± 7.6	84.6 ± 11.0	84.6 ± 11.0
MIXEDSTABLE + persistent color rand.	95.1 ± 0.4	98.0 ± 0.2	95.0 ± 0.3	92.9 ± 0.7	92.9 ± 0.7

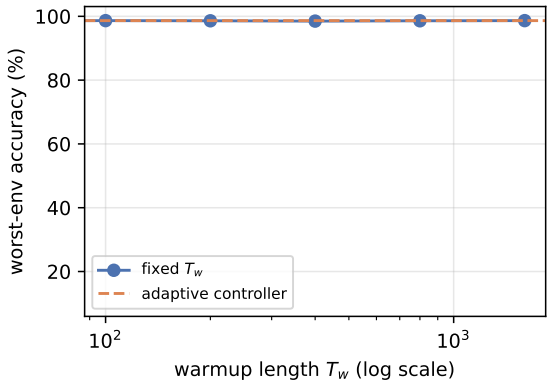

Figure 1: Worst-environment accuracy vs. warmup length T_w (log scale), 100%-label protocol (plain GIN + linear head; 3,000 graphs; 3 seeds), *best-IID-validation* read-out. A read-out audit (App. D.3) shows at T_w=100 the selected checkpoint falls inside the warmup phase (final-epoch accuracy 92–94 vs. selected 99), while at $T_w \geq 400$ the two read-outs coincide.

720 total steps; final-epoch read-out). Result (App. D.3): release-point accuracy is 98.6–99.2 in all six forks (warmup success 6/6), but preservation is observed in a subset of runs and is not reliable under the current protocol: several arms—including continued randomization—drop to chance 33.4 in most seeds, while the GIN release-without-consistency arm retains 72–93. The collapse is not confined to released arms: the persistent arm also collapses in most seeds, and a read-out audit attributes the earlier "saturation" claim for $T_w \geq 100$ to best-validation selection (the selected checkpoint sits inside the warmup; at $T_w \geq 400$ final-epoch accuracy is stable at 99.6). High accuracy during the intervention does not imply reliable preservation after extended training.

5.4 Q3: WHAT DO THE ADDITIONAL COMPONENTS CONTRIBUTE?

Consistency channels. At 10 seeds the struct-only and dual-channel variants are statistically indistinguishable (94.4±5.6 vs 93.6±4.2, mean diff −0.8, 95% CI [−5.1, +3.5]; Welch p=0.72): under this low-label protocol the intervention spans the entire budget, so the consistency terms operate alongside ongoing shortcut destruction, not after it. Externally the question splits: on GOOD+Color the machinery collapses to chance in every arm while the prior-only control holds 62%, and on CMNIST-v3 the dual channel shows the only margin over its single-channel variant and the prior-only control (38.6±1.1 vs. 29.3±3.0 / 31.8±8.4; App. D.8). A module-level ablation completes

Table 3: External benchmark variants (worst-environment accuracy, 3 seeds unless noted; official datasets vs. our variants marked; full arm-by-arm data in App. D.8). "steps" lists T_{total}/T_w for the GIN+intervention arm; checkpoint = final epoch except GOOD-Motif/v1 (best-val, historical single runs, no seed replication—marked †). MIXEDSTABLE column shows its best arm.

Benchmark	type	K/ρ	steps	ERM	GIN+interv.
GOOD-Motif†	official	3 / 100%	150/100	100.0	100.0
GOOD-CMNIST (v1)†	official, 4×4-grid reimpl.	3 / 100%	150/100	54.5	56.2
GOOD-CMNIST (v3)	our variant	5 / 100%	560/80	20.1 ± 2.8	31.8 ± 8.4
GOOD+Color	our hybrid	3 / 100%	200/100	56.8 ± 3.5	62.0 ± 8.3
REDHOUSE	ours	3 / 10%	120: 100 (GIN) / all (MIXEDSTABLE)	14.2 ± 2.1	96.1 ± 2.2

the picture (App. D.9): removing the attention gate substantially reduces performance within this configuration (41.5 ± 19.6), while for the other modules we do not establish individual contributions.

Intervention validity. The color-histogram probe drops from 100.0% to chance ($33.8 \pm 4.1\%$) under the intervention while the degree-histogram probe retains 83.5% (Sec. 3.1, App. D.4): the intervention substantially reduces label predictability from the evaluated color probe, and the structure probe survives.

Training-side bias composition. A 2×2 grid over training biases $(b_s, b_f) \in \{1/3, 0.9\}^2$ (fixed test envs, 10% labels, 120 updates, final-epoch read-out, 100 test graphs per environment vs. the main tables' 1,500; App. D.5): feature bias is sufficient to induce ERM failure in this experiment (ERM 7.7–9.0 whenever $b_f{=}0.9$, regardless of b_s). After targeted randomization, however, the doubly biased setting retains a larger performance deficit: adding structural bias costs -4.7 points at $b_f{=}1/3$ ($98.0 \to 93.3$) but -11.0 at $b_f{=}0.9$ ($99.3 \to 88.3$), a difference-in-differences point estimate of -6.3 points. Whether this reflects a reproducible interaction requires a dedicated per-seed analysis; we report the point estimate without claiming an interaction effect.

Randomizing a block containing redundant label information. Injecting class-indicating signal into the intervened block (motif-node colors carry the class color with probability α; base-node colors remain the shifting nuisance; App. D.6): persistent randomization shows no clear degradation across $\alpha \in [0, 1]$ (87.0 ± 2.6 at $\alpha{=}0$ vs. 87.3 ± 3.5 at $\alpha{=}1$), while ERM improves with α (7.7 to 18.7) as the block becomes informative. Because the motif structure fully determines the label in this benchmark, the injected signal is *redundant*: this experiment shows that deleting redundant signal from the intervened block costs little here; it does *not* establish the failure boundary for an incorrect prior that removes irreplaceable information. The worst-environment gap between the arms narrows from 79 to 69 points as ERM increasingly exploits the block.

Label efficiency. The sparse stable-variable learner fails on feature-shifted environments at *every* label fraction (2%–100%, worst 10–14%; App. D.2): label quantity is not the bottleneck in this setting.

5.5 EXTERNAL VARIANTS

The benchmark variants exhibit heterogeneous intervention gains (Table 3). **GOOD-Motif**: with only basis covariate shift, all methods reach 100%. **GOOD-CMNIST**: a three-round calibration study (App. D.10) yields two community-level lessons: (i) the calibration trap—with off-class colors sampled uniformly from *all* K colors (the class color included), a generation parameter ρ yields an effective test correlation $\rho + (1 - \rho)/K$; the v1 setting used $\rho_{\text{test}}{=}1/9$, $K{=}3$, giving $11/27 \approx 0.41$ where near-zero was intended; (ii) MNIST superpixels have large γ_s (digit shape readable from pixel means), making them resistant to feature-shortcut dominance. **GOOD+Color**: our hybrid with injectable color block shows a moderate intervention gain ($+5.2$) and a failure of the full framework (all machinery arms at chance 33.5). On CMNIST-v3 the dual-channel arm shows the only machinery margin in our study (38.6 vs. 29.3 single-channel, 31.8 prior-only). These outcomes do not establish that mixed shifts necessarily cause stronger shortcut dependence than feature shifts alone.

6 CONCLUSION

We studied targeted feature interventions under controlled graph distribution shifts. On RED-HOUSE, the observed robustness gains are largely explained by randomization of a known nuisance block, and the evaluated composite framework does not show an established advantage over a prior-matched GIN with the finite-warmup intervention (95.7 ± 2.6 vs. 93.6 ± 4.2 at $n{=}10$, $p{=}0.20$; stacking IRM on the intervention is significantly harmful, -14.5 points, $p{<}10^{-3}$). Persistent randomization and finite warmup should be distinguished: high accuracy during intervention does not imply reliable preservation after the original shortcut is restored (a same-checkpoint fork isolates the effect of the returning shortcut with matched training trajectories; App. D.3). External benchmark variants further reveal heterogeneous gains and failures of the full framework (GOOD+Color: all machinery arms at chance; CMNIST-v3: a moderate dual-channel margin, App. D.8). These findings support matching prior information and evaluation protocols when assessing graph robustness methods. The effects of imperfect prior knowledge, structure–feature interactions, and long-term post-release training remain important boundaries to establish.

REFERENCES

Anonymous. Out-of-distribution generalization in graph foundation models. *arXiv preprint arXiv:2601.21067*, 2026a.

Anonymous. When do graphs learn shortcuts? gradient dominance, phase transition, and a warmup intervention. *Under review (same conference, companion submission)*, 2026b. Anonymous dual submission; cited for coherence, not relied upon for any proof.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. 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.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning (ICML)*, 2021.

Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4:110–115, 2022.

Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, et al. Generalizing graph neural networks on out-of-distribution graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. arXiv:2111.10657.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. *Algorithmic Learning Theory (ALT)*, 2005.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2022.

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.

Haoyang Li, Yang Wang, et al. Disentangled graph self-supervised learning for out-of-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2024.

Haoyang Li, Xin Wang, Xueling Zhu, Weigao Wen, et al. Disentangling invariant subgraph via variance contrastive estimation under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2025.

Tianyin Liao, Ge Lan, Rui Chen, Ran Zhang, Zhiming Chen, Xiao Wang, and Ziwei Zhang. Unifying graph out-of-distribution generalization and detection through spectral contrastive invariant learning. In *Proceedings of the ACM Web Conference (WWW)*, 2026.

Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. ZIN: When and how to learn invariance by environment inference? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bowen Lu, Lianqiang Yang, Teng Li, et al. Causal-guided representation learning for node-level out-of-distribution generalization. *arXiv preprint arXiv:2603.24304*, 2026.

Yihong Luo, Yuhan Chen, Siya Qiu, Yiwei Wang, Chen Zhang, Yan Zhou, Xiaochun Cao, and Jing Tang. Fast graph sharpness-aware minimization for enhancing and accelerating few-shot node classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Yang Qiu, Yixiong Zou, et al. Quantifying distributional invariance in causal subgraph for IRM-free graph generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.20295.

Yang Qiu, Yixiong Zou, and Jun Wang. Rethinking graph generalization through the lens of sharpness-aware minimization. *arXiv preprint arXiv:2602.08855*, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Zheyan Shen, Peng Cui, Kun Kuang, Bochen Li, and Pengtao Chen. Causally regularized learning with agnostic data selection bias. In *ACM Multimedia (MM)*, 2018.

Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bochen Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2020.

Yili Wang, Kaixiong Zhou, Ninghao Liu, Ying Wang, and Xin Wang. Efficient sharpness-aware minimization for molecular graph transformer models. *arXiv preprint arXiv:2406.13137*, 2024. Distinct method sharing the name; see nomenclature note in Sec. 2.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations (ICLR)*, 2022a. arXiv:2202.02466.

Qitian Wu, Fan Nie, Chenxiao Yang, and Tianyi Bao. Graph out-of-distribution generalization via causal intervention. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.11494.

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)*, 2022b.

Renzhe Xu, Peng Cui, Zheyan Shen, Xingxuan Zhang, and Tong Zhang. Why stable learning works? A theory of covariate shift generalization. Preprint, 2021.

Renzhe Xu, Zheyan Shen, Xingxuan Zhang, Tong Zhang, and Peng Cui. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *International Conference on Machine Learning (ICML)*, 2022.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen, and Haoxin Liu. Deep stable learning for out-of-distribution generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yuxiang Zhang, Enyan Dai, et al. SCL-GNN: Towards generalizable graph neural networks via spurious correlation learning. *arXiv preprint arXiv:2603.08270*, 2026.

A Appendix: Theoretical Background

The formal statements that motivated earlier versions of the method stack are recorded here as background facts. The main text makes no recovery, identification, or invariance claim that depends on them.

A.1 Risk invariance under stable-label mechanisms (background)

Lemma 1 (Risk invariance; background fact)**.** *Assume (i)* label sufficiency*: $Y \perp Z_S \mid Z_C$; (ii) the test shift acts only on the unstable block: $P_{\mathrm{te}}(Z_C) = P_{\mathrm{tr}}(Z_C)$ and $P_{\mathrm{te}}(Y \mid Z_C) = P_{\mathrm{tr}}(Y \mid Z_C)$; (iii) g is Z_C-measurable. Then $R_{\mathrm{te}}(g) = R_{\mathrm{tr}}(g)$.*

Proof. For g measurable in Z_C, $R_{\mathrm{te}}(g) = \mathbb{E}_{P_{\mathrm{te}}(Z_C)}\mathbb{E}_{P_{\mathrm{te}}(Y|Z_C)}[\ell(g(Z_C), Y)]$; both factors equal their training counterparts by (ii). □

Scope. The proof uses only the invariance of $P(Z_C)$ and $P(Y \mid Z_C)$; assumption (i) is a property of the label mechanism ($Y = C + S + \epsilon$ with an S-shift is the standard counterexample to stronger readings).

A.2 What pairwise HSIC monitoring does and does not establish

Pairwise dependence diagnostics can detect violations of cross-block independence, but small pairwise HSIC values do not establish block independence or identify label-sufficient coordinates:

- *Detection direction.* Under block independence $Z_C \perp Z_S$, every cross pair ($i \in C, j \in S$) has population $\mathrm{HSIC}(Z_i, Z_j) = 0$, and the empirical RFF estimate concentrates at $O_p(D^{-1/2} + N^{-1/2})$ (Rahimi & Recht, 2007; Gretton et al., 2005); a candidate split whose cross-pair estimates clearly exceed this slack is inconsistent with block independence. We do not state a threshold or multiple-testing guarantee.
- *No identification (XOR counterexample).* Let B_1, B_2 be independent fair bits and $Z_1 = B_1$, $Z_2 = B_2$, $Z_3 = B_1 \oplus B_2$: each pair is independent, yet any two coordinates determine the third, so *every* bipartition of $\{1, 2, 3\}$ violates block independence and every bipartition passes the pairwise cross test simultaneously—passing cannot single out a split. A shared latent factor does not imply observable pairwise dependence; block-level dependence measures would be required under such mixing.
- *No risk bound.* No analysis links small empirical HSIC values to the risk of predictors built on approximately decorrelated coordinates; the XOR instance shows such a bound cannot hold unconditionally.

A.3 Counterexample for covariance-based stable-set selection

The following counterexample concerns an *idealized* selector—among all $\binom{m}{k}$ candidate sets, choose the one minimizing the maximum absolute cross-boundary covariance—and shows this population objective can prefer an unstable set. It does not assert that any particular optimization algorithm must fail, nor does it establish a converse.

Let $U, \epsilon_2, \epsilon_3, \epsilon_Y$ be independent standard Gaussians and

$$Z_1 = U, \qquad Z_2 = U + \epsilon_2, \qquad Z_3 = 0.1\,U + \epsilon_3, \qquad Y = U + \epsilon_Y.$$

The joint law is strictly positive, and the unique minimal stable set of size 1 is $C = \{1\}$. With $\mathrm{Cov}(Z_1, Z_2) = 1$, $\mathrm{Cov}(Z_1, Z_3) = \mathrm{Cov}(Z_2, Z_3) = 0.1$, the min–max objective $\Delta(S) = \max_{i \in S, j \notin S} |\mathrm{Cov}(Z_i, Z_j)|$ gives $\Delta(\{1\}) = \Delta(\{2\}) = 1$ but $\Delta(\{3\}) = 0.1$: the selector returns $\{3\}$, which is not stable, while the stable set $\{1\}$ is penalized by its covariance with the unstable copy Z_2. The failure occurs with population covariances known exactly. □

Why it fails. Stability is label-conditional; the objective is label-blind. A stable coordinate covaries with unstable ones exactly when the unstable ones copy the signal, so minimizing cross-boundary covariance rewards discarding signal-coupled coordinates. The hard-concrete mask and reweighter

in MIXEDSTABLE are retained as capacity-control heuristics; the counting comparison ($\binom{m}{k}$ candidates vs. $\Theta(m^2)$ pairs) counts test statistics, not samples, and no sample-complexity theory of stable-set selection is claimed.

A.4 REFERENCE: ASSUMPTIONS UNDER WHICH AN ANCHORED-CONSISTENCY STATEMENT WOULD HOLD

For reference we record the assumptions under which an invariance statement for the anchored consistency objective could be proved. The implemented operators (uniform edge dropout; cross-node color-row permutation) are *not* shown to satisfy them, and the main text relies on the empirical evaluation instead. A formal statement would require at minimum:

(a) each augmentation family changes only its own channel's nuisance conditional and leaves the stable representation invariant along orbits;

(b) label-preserving transformations: $Y \perp \mathcal{T} \mid Z_C$;

(c) anchoring: supervised CE on the original and both augmented views with $\lambda_{\text{cons}} > 0$ (the bare two-view KL admits degenerate constant solutions);

(d) adequate coverage: with Q the distribution of augmented views induced by the training augmentation process, the target interventional distribution satisfies $P_{\text{target}} \ll Q$ (absolute continuity with respect to a measure, not with respect to a set);

(e) realizability: the hypothesis class contains a predictor whose law equals the Bayes law $p^*(Y \mid Z_C)$;

(f) compatibility: the sparsity penalties vanish at that predictor—and simultaneously so with a nontrivial Bayes predictor, which is not automatic.

We deliberately do *not* state a proposition from this list. Connecting the assumptions to a conclusion would additionally require controlling the gap between "predictive law agrees almost surely" and "risk values are equal" on the augmented support, verifying that the minibatch color permutation's induced Q covers the claimed target family, and matching the implemented composite objective (which adds reweighting, attention, and ℓ_0 terms) to the idealized anchored objective—none of which we carry out. On REDHOUSE-like data, (b) is plausible for the color permutation (colors are label-independent given the motif) and (e) is plausible when the label is a noiseless function of the motif and the encoder class is sufficient; the implemented edge dropout satisfies (b) only approximately (it can perturb motif connectivity at finite drop rates). These plausibility remarks do not establish coverage or compatibility; the robustness of the stack is assessed empirically (Sec. 5).

A.5 ALGORITHM DETAILS

MIXEDSTABLE training step. Each training step: (1) encode and compute attention-masked readout z_{G_c}; (2) supervised CE with reweighter weights; (3) consistency between structure- and feature-augmented views (prediction-level symmetric KL); (4) DIR-consistency InfoNCE on z_{G_c}; (5) sparsity penalties; every T steps, (6) reweighter update with decorrelation loss on detached z_{G_c}. Full hyperparameters in App. B.1.

B APPENDIX: EXPERIMENTAL DETAILS

B.1 HYPERPARAMETERS

B.2 COMPUTE AND SEED POLICY

Compute: single RTX 5090 (24GB); full study approximately one GPU-day. Seed counts differ by experiment and are stated per table: main tables and controls use 3 seeds; the statistical study (App. D.7) uses n=10; the read-out audit (App. D.3, E56) uses 2; GOOD-Motif and CMNIST-v1 rows in Table 3 are historical single runs (marked †).

Table 4: Default hyperparameters.

Module	Hyperparameter	Value
GRAPHSAM	feature attack radius ϵ	0.05
	PGD steps	1
	weight radius ρ	0.02
	projector/predictor hidden	128
	SSL form	BYOL (symmetric, stop-grad)
HSIC-SSL	λ_{HSIC}	1.0
	Fourier features D	32
	kernel bandwidth σ	1.0; resample every 10 epochs
SPARSESTABLE	ℓ_0 penalty λ_{ℓ_0}	0.05
	alternation period	5 steps
MIXEDSTABLE	λ_{cons}	1.0
	λ_{dir}	1.0
	attention sparsity λ_a	0.01
	edge-drop rate (structure env)	0.2
	block-permutation rate (feature env)	1.0 (full permutation)
	InfoNCE τ	0.2
backbone		GIN, 3 layers, hidden 64, sum pool
optimizer		Adam, lr 10^{-3}, weight decay 10^{-4}
batch / epochs		256 / 50 (SSL), 60 (stable)

B.3 PER-TABLE PROTOCOL PROVENANCE

Each main-text table corresponds to a distinct script/protocol; we list the mapping explicitly (all results keyed by JSON in the code release): SSL probes (Spurious-Motif, linear probe, $5\%/100\%$ labels, 50+probe epochs): `run_exp1.py`. Table 1 (mixed-shift main): `run_exp2.py`, REDHOUSE, ρ=10% (300 labeled graphs), 60 epochs, batch 256, MixedStable defaults (App. B.1). Consistency-channel ablation: `run_ablation.py`, same protocol as Table 1; the warmup phase spans the entire training budget there (default T_w=800 steps exceeds the $\sim$120 optimizer steps of the 10%-label protocol), i.e. the color shortcut is destroyed in the original view for the whole run; the persistent- randomization-without-consistency row is MIXEDSTABLE with λ_{cons}=0 (attention, mask, reweighting, DIR and worst-of-pseudo-env CE remain), *not* plain ERM plus the intervention (that control is Table 2). Table 3 and Fig. 1 (T_w sweep): `run_warmup_schedule.py`, 100% **labels** (3,000 graphs), plain GIN + linear head, so its 98.6% is not comparable to the 10%-label tables (89.7–94.5); label-percentage protocols are stated per table. Appendix warmup-sensitivity sweeps that override defaults (λ_{dir}, $RFF\ D$, bandwidth) are marked per entry.

B.4 DATASET STATISTICS

REDHOUSE (default b_{struct}=b_{feat}=0.9): 3,000 training graphs (ρ=10% labeled $\Rightarrow$ 300 labeled), 300 validation, $4 \times 1{,}500$ test; 10–27 nodes per graph (base 10–20 + motif 5/5/6 for house/crane/cycle); motif nodes are the last 5–6 indices; feature dimension 10 (degree one-hot 7 $\oplus$ color one-hot 3); 3 classes. Spurious-Motif: 9,000 train / 3,000 test, bias 0.9. *Checkpoint rules.* The low-label main tables (Tables 1, 2) use final-epoch read-outs; the warmup-schedule sweep (Fig. 1) uses best-IID-validation selection (see the read-out audit in App. D.3; the two read-outs are reported separately wherever they differ).

C CHANNEL-SENSITIVITY PROFILING OF PRETRAINED ENCODERS

The $(b_{\text{struct}}, b_{\text{feat}})$ grid doubles as a diagnostic for pretrained encoders—the evaluation axis that the GFM-OOD survey (Anonymous, 2026a) identifies as missing. Protocol: pretrain on a multi-domain pool (four (b_s, b_f) combinations) with standard SSL, then linear-probe each of the 25 grid cells.

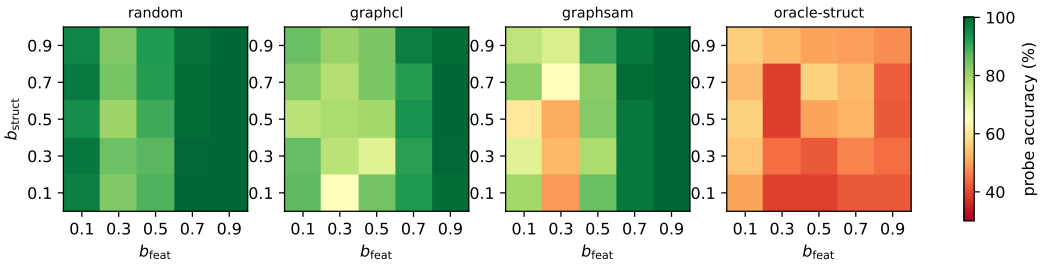

Figure 2: Probe accuracy over the grid for four pretraining variants. SSL introduces asymmetric channel biases; RFF-HSIC decorrelation flattens the asymmetry at an availability cost.

Findings. (i) A random encoder is nearly channel-balanced (93.6% grid mean; sensitivities $3.9/1.4$)—the bias is introduced by training, not architecture. (ii) GraphCL and adversarial+weight-perturbed SSL *amplify* structure-channel sensitivity (16.2 and 26.7) and lower the grid mean ($86.3/82.5$): self-supervised objectives take sides in the channel race. (iii) Adding an RFF-HSIC dimension-decorrelation regularizer restores balance (structure sensitivity 2.7) but collapses probe availability at $\lambda{=}1$ (39–43% mean); a unified-protocol sweep locates a small-λ optimum (0.05, $+6.7$ points) with the trade-off non-monotone in λ and robust qualitatively across two protocols. These profiles constitute, to our knowledge, the first per-channel sensitivity measurement of graph pretraining, and the natural first evaluation for graph foundation models claiming channel robustness. The profiles are *static* (snapshot-based); the companion theory (Anonymous, 2026b) additionally identifies a complementary *trajectory-only* label-free statistic (per-feature-group flip-jitter asymmetry), which cannot be recovered from snapshots.

D Additional Experimental Details

D.1 SSL collapse diagnostics

As stated in Sec. 4.1, we monitor representation variance and effective rank $\mathrm{effrank}(\Sigma) = \exp(-\sum_i p_i \log p_i)$, $p_i = \lambda_i / \sum_j \lambda_j$, on a fixed 512-graph probe batch during HSIC-SSL training (synthetic Spurious-Motif-style data, 3 seeds, 30 epochs). Two modes: *bare* (plain alignment loss, no predictor/stop-grad) vs. the full *BYOL recipe* (asymmetric predictor + stop-grad). Results at the final epoch (mean over seeds): bare variance $1{\times}10^{-4}$ ($\approx$ collapsed; alignment loss $\to 0.0005$), BYOL-recipe variance $0.26{\pm}0.05$ with loss stabilized at $0.68{\pm}0.09$—the recipe prevents the variance collapse the bare objective drives toward. *Honest note on rank*: effective rank stays low in *both* modes (1.7 vs. 3.3 of 128 projected dims): the recipe preserves variance but does not by itself guarantee high-dimensional spread; spectral diversity of this SSL family is not established by these diagnostics and remains an open property. These diagnostics bound what we claim for the objective.

D.2 Label-efficiency sweep

Worst-environment accuracy of the sparse stable-variable learner (MixedStable minus warmup/consistency) vs. label fraction on RedHouse: 2%: 12.1, 5%: 13.4, 10%: 13.9, 30%: 10.5, 100%: 13.5. The failure is invariant to supervision quantity.

D.3 Post-release same-checkpoint fork (E53) and the earlier from-scratch control (E49)

E53 (same-checkpoint fork; primary). 100% labels, $3{,}000$ graphs, batch 256, 720 steps. All arms share an *identical* warmup trajectory (persistent color randomization for $T_w{=}100$ steps, plain GIN or full MixedStable); the model and optimizer state are then snapshotted and each arm forks from the same checkpoint: *rel-nocons* (original view returns, no extra regularizer), *rel-cons* (original view returns; worst-of-pseudo-env CE + symmetric-KL consistency added), *persistent* (randomization continues). Final-epoch read-out; warmup success rate $6/6$ (release-point worst-env 98.6–99.2). Per-seed final worst-env:

Fork	GIN finals	MS finals
rel-nocons	93.0/92.2/72.2	99.6/99.4/33.4
rel-cons	99.4/33.4/33.4	99.4/33.4/33.4
persistent	99.8/33.4/33.4	99.8/33.4/33.4

Preservation is observed in a subset of runs and is not reliable under the current protocol—for either the consistency regularizer or continued randomization. The collapse is *not* caused by release alone: even the persistent arm, which never sees the original colors, collapses in most seeds after forking, indicating that continued training from a T_w=100 checkpoint is itself unstable at this budget. High accuracy at the release point does not imply reliable preservation after extended training; the earlier saturation claim for $T_w \geq 100$ under this protocol was a best-validation selection artifact (the validation split shares the training shortcut correlation; see the read-out audit below).

E49 (earlier from-scratch control; superseded by design). Arms trained from scratch with different seeds confound warmup quality with release dynamics; its readings (re-collapse under the full objective without consistency; plain-GIN arm most robust) partially agree with E53, but E53's fork data do not support a causal reading of consistency removal—the MS release-without-consistency arm retains 99% in two of three seeds—so we state only that preservation is unreliable in both designs. The fork design is the cleaner comparison. Data: ms-cons finals 99.6/33.4/33.4; ms-nocons 33.4×3 (from successful warmups 98.6→33.4); plain-GIN 89.5/80.9/73.1.

Resume-consistency audit and consistency decomposition (E57/E58). Two follow-ups tighten the fork analysis. *E57*: training 720 steps continuously (persistent randomization) matches, per seed and exactly, training 100 steps followed by save/restore of the full state (model, optimizer, RNG) and 620 further steps (99.8/99.8/100.0 both ways, 3/3 seeds)—the save/restore operation itself does not perturb results, so the E53 collapses are optimization phenomena, not checkpointing artifacts. Notably, the E57 continuous persistent arm does *not* collapse (3/3 at ≥99.8), differing from E53's persistent fork only in the permutation-RNG stream: continued randomization is stable under some random streams and collapses under others, which is the operational meaning of "seed instability" here. *E58*: forked arms with the augmented-view CE always on and only the symmetric-KL term toggled show CE-only arms retaining accuracy in 3/3 seeds (99.6/99.6/99.0) while adding KL collapses in 2/3 (33.4/33.4/99.4)—the instability observed after release is attributable to the consistency (KL) term rather than the augmented CE, at this seed count.

Read-out audit (E56). The warmup-schedule sweep (Fig. 1) originally reported 98.5–98.6% for all $T_w \geq 100$ under *best-IID-validation* checkpoint selection. Re-running the same schedule with both read-outs (3,000 graphs, 60 epochs, 2 seeds): at T_w=100, final-epoch worst-env is 94.0/91.6 vs. best-val 98.6/99.4—and the best-val checkpoint falls at epoch 5–6, *inside* the warmup phase, so the selection effectively snapshots the warmup-end model and masks the post-warmup drift; at T_w=400 the two read-outs coincide (99.6). The IID validation split retains the training shortcut correlation and cannot detect post-release decay. Final-epoch and selected accuracies are therefore reported as separate read-outs wherever they differ; Fig. 1 shows the selected read-out.

Warmup transition band (selected read-out). Under the 100%-label, plain-GIN protocol of Fig. 1 (3 seeds, best-val read-out), intermediate warmup lengths resolve the transition between shortcut and structural occupation (worst-environment accuracy): T_w=5: 16.0±8.7; T_w=10: 17.6±5.7; T_w=25: 31.8±18.8; T_w=50: 72.9±20.9; T_w=75: 94.1±6.7; T_w=0 gives 10.4%. The jump concentrates between T_w=25 and T_w=75 optimizer steps (≈ 2.1–6.3 epochs at batch 256), with large seed variance inside the band—consistent with threshold-like occupation dynamics (Anonymous, 2026b). We emphasize these are *selected* accuracies; E53/E56 show final-epoch accuracy from the same T_w=100 warmups decays in most seeds when training continues to 720 steps.

D.4 Intervention validity checks (E52)

Color-histogram probes (ridge one-vs-all on mean color one-hots, 100 random 60/40 splits) on the 300-graph low-label training pool: original features 100.0 ± 0.0 → after one pooled-node permutation draw 33.8±4.1 (chance 1/3). Degree-histogram probe: 83.5±3.2; motif-restricted degree probe (ground-truth mask, evaluation-only): 100.0; graph-size probe: 36.4 ± 3.4. A mean-pooling MLP

(no message passing) trained on original vs. permuted features gives identical accuracies (worst-env 0.08 both): without message passing the task is not solvable from these features regardless of the intervention, i.e. the intervention destroys no information the no-MP model could use. Together: the intervention removes color–label information, and the structure information that GINs can exploit survives.

D.5 Training-side 2×2 bias grid (E54)

$(b_s, b_f) \in \{1/3, 0.9\}^2$ training cells, fixed standard test environments, ρ=10% labels, 120 updates, final-epoch read-out, worst-env mean ± std over 3 seeds:

Training cell (b_s, b_f)	ERM	GIN + persistent rand.
(1/3, 1/3)	93.7 ± 4.6	98.0 ± 1.0
(0.9, 1/3)	92.0 ± 2.6	93.3 ± 3.5
(1/3, 0.9)	9.0 ± 2.6	99.3 ± 0.6
(0.9, 0.9)	7.7 ± 2.1	88.3 ± 2.5

ERM collapses exactly when b_f=0.9, regardless of b_s: feature bias is sufficient to induce ERM failure in this experiment. After targeted randomization the doubly biased setting retains a larger deficit (-4.7 points when adding b_s at b_f=1/3; -11.0 points at b_f=0.9; difference-in-differences point estimate -6.3). Whether this reflects a reproducible interaction requires a dedicated per-seed analysis; we report the point estimate only. Note this grid uses 100 test graphs per environment (freshly generated draw), versus 1,500 in the main tables, so absolute values are not directly comparable (e.g. ERM $(0.9, 0.9)$: 7.7 here vs. 14.2 in Table 1).

D.6 Randomizing a block containing redundant label information (E55)

Motif-node colors carry the class color with probability α (injected signal inside the intervened block); base-node colors remain the b_{feat}-correlated nuisance. ρ=10%, 120 updates, 3 seeds, worst-env:

α	0	0.25	0.5	1.0
GIN + persistent rand.	87.0 ± 2.6	87.7 ± 2.1	87.3 ± 2.5	87.3 ± 3.5
ERM	7.7 ± 2.1	10.0 ± 3.5	11.7 ± 5.5	18.7 ± 5.5

No clear degradation is observed across α: at α=1 (every motif node colored by class) randomizing the block costs at most $\sim$0 points relative to α=0. Because the motif structure fully determines the label, the injected signal is redundant, and ERM's failure mode remains dominated by the base-node nuisance; ERM improves with α as the block becomes informative. The worst-env gap between the arms narrows from 79 to 69 points. Scope: this experiment shows deleting *redundant* signal from the intervened block costs little when the structure channel is sufficient; it does not test the failure boundary for an incorrect prior that removes irreplaceable information—on datasets where the intervened block is the *only* predictive channel the same randomization would destroy the task.

D.7 Ten-seed statistical study (E50)

Key comparisons re-run at n=10 seeds under the exact Table 1 protocol (ρ=10%, 60 epochs; finite-warmup arms use T_w=100), worst-environment accuracy (per-seed minimum first, then aggregated). Intervals are Welch two-sample 95% CIs for the mean difference; paired 95% CIs (same data split and seeds) are given beneath and agree in all three conclusions:

Pair	means (worst-env)	Welch 95% CI of diff	p (Welch/Wilcoxon)
warmup vs. full	95.7±2.6 vs 93.6±4.2	$[-1.2, +5.4]$	0.20 / 0.32
paired		$[-1.6, +5.8]$	0.23
warmup vs. IRM+warmup	95.7±2.6 vs 81.2±7.5	$[+9.0, +20.0]$	$< 10^{-3}$ / 0.002
paired		$[+10.3, +18.6]$	$< 10^{-3}$
full (dual) vs. struct-only	93.6±4.2 vs 94.4±5.6	$[-5.5, +3.9]$	0.72 / 0.70
paired		$[-5.1, +3.5]$	0.68

The headline comparison does not establish an advantage of the full stack over the finite-warmup control: the interval contains zero and the point estimate favors the simple control (nonsignificance $\neq$ equivalence). The one strong effect concerns the IRM arm: with its penalty active from step 0 (n=10), IRM on top of the intervention is significantly harmful (-14.5 points). An earlier 3-seed run of the same arm (96.2 ± 2.6) used a step-50 activation buffer that left the penalty mostly inactive; we treat the always-on behavior as IRM's and flag the discrepancy. The dual-vs-struct-only gap remains within noise on REDHOUSE; external single-/dual-channel data appear in App. D.8.

D.8 EXTERNAL SINGLE- VS. DUAL-CHANNEL CONTROLS (E51)

Single- vs. dual-channel consistency on the two external benchmarks, with machinery and plain warmup receiving the *same* block prior (per-dataset color-block permutation). Worst-environment accuracy, 3 seeds (GOOD+Color: 4 shifted test envs; CMNIST-v3: ρ_{test}=0 OOD, chance = $1/5$):

Arm	GOOD+Color	CMNIST-v3 (OOD)
ERM	56.8 ± 3.5	20.1 ± 2.8
Warmup (prior only)	62.0 ± 8.3	31.8 ± 8.4
ms-nocons (λ_{cons}=0)	33.5 ± 0.0	31.4 ± 3.6
ms-struct (single channel)	33.5 ± 0.0	29.3 ± 3.0
ms-dual (dual channel)	33.5 ± 0.0	$\mathbf{38.6}\pm1.1$

Two findings. (i) On GOOD+Color the MIXEDSTABLE machinery *collapses to the class-count chance level* (33.5%, K=3) in every arm and seed, while the plain warmup prior alone holds 62.0 ± 8.3—the machinery is not merely unhelpful but destructive on this benchmark, and the single-/dual-channel question is moot there (both at chance). (ii) On CMNIST-v3 the dual-channel arm reaches 38.6 ± 1.1 vs. single-channel 29.3 ± 3.0 and prior-only 31.8 ± 8.4—the *only* setting in our study where the dual-channel consistency shows a margin over both the single-channel variant and the prior-only control (roughly +7–9 points at this seed count), directionally consistent with the mixed-shift motivation of the dual-channel design, though we note the seeds are few and the machinery arm without consistency sits between them (31.4 ± 3.6). Read together with E50: on REDHOUSE the machinery adds nothing over the prior (p=0.72 for dual vs. struct); externally it either collapses (GOOD+Color) or shows a moderate dual-channel margin (CMNIST-v3)—the machinery's value, where it exists at all, is benchmark-conditional and does not survive the prior-matched control on the benchmark it was designed for.

D.9 MODULE ABLATION OF MIXEDSTABLE

Removing one module at a time from full MIXEDSTABLE (Table 1 protocol, 3 seeds, worst-environment accuracy): full re-run 93.7 ± 3.2; without attention gating 41.5 ± 19.6 (a large drop in this configuration); without the ℓ_0 mask penalty 94.5 ± 2.5; without DIR consistency 93.9 ± 1.8; without sample reweighting 91.3 ± 7.8. Within this configuration, removing the attention gate substantially reduces performance while the mask, DIR, and reweighting modules each contribute within one seed-level standard deviation. We do not localize gradients to motif nodes and offer no mechanistic explanation for the attention gate's observed importance; read together with Table 2 (where plain ERM plus the intervention alone matches the full method), the prior carries the benchmark, and this ablation characterizes only the composite framework's internal sensitivities.

D.10 GOOD-CMNIST CALIBRATION ROUNDS

v1 (3 classes, 4×4 grid, ρ_{test}=1/9): mild +1.7 gain; the calibration trap—off-class colors sampled uniformly from all K colors including the class color—yields an effective test correlation $\rho + (1-\rho)/K = 11/27 \approx 0.41$ instead of near-zero. *v2* (10 classes, ρ_{test}=0): both methods collapse to chance—the structure task exceeds a 3-layer GIN's capacity ($\gamma_s \approx 0$); warmup cannot substitute for structural capacity. *v3* (5 classes, 5×5, ρ_{test}=0): in the historical single-seed runs the warmup gave +1.5 OOD; the replicated 3-seed protocol of App. D.8 gives +11.7 (31.8 ± 8.4 vs. 20.1 ± 2.8), with a zeroed-color probe gain of +4.8 points ($15.7 \to 20.5\%$)—the structural channel benefits even when the shortcut cannot fully dominate.

D.11 HYPERPARAMETERS

All methods share the GIN backbone (3 layers, hidden 64, sum pooling), Adam with lr $\in [10^{-4}, 10^{-3}]$ tuned per method on validation IID accuracy, ϵ=0.05, ρ=0.02 (SAM parameters), batch 256, 60 epochs (supervised) / 50 epochs (SSL probes). MIXEDSTABLE uses the defaults of App. B.1 in all main tables (λ_{cons}=1, λ_{dir}=1, λ_a=0.01); warmup T_w=800 by default, which under the 10%-label protocol ($\sim$120 steps) covers the whole run (App. B.3); the warmup-sensitivity sweep below overrode λ_{dir}=0.1, λ_a=10^{-4}, D=256, median-heuristic bandwidth, resampling every 5 epochs, and adaptive-controller settings (EMA β=0.99, release threshold 2$\times$ structural EMA). RFF: D=32 elsewhere.

E REPRODUCIBILITY STATEMENT

All experiments use a fixed codebase (data generators, method implementations, training loops, and audit probes) released with the supplementary material. Data generation is fully seeded: every benchmark split is a deterministic function of the seed. Seed counts are stated per experiment (main tables and controls: 3 seeds; statistical study: n=10; read-out audit: 2; two historical external rows are single runs and marked as such). Complete hyperparameters are listed in App. D.11; result JSONs for every table and figure are included in `results/`, keyed by table/figure number.

F AI USE STATEMENT

The authors used AI assistants (LLM-based coding and writing aids) for language polishing and boilerplate code scaffolding under direct author supervision. All scientific content—problem formulation, method design, theoretical results, experimental protocols, and the analysis and interpretation of results—was produced and verified by the authors, who take full responsibility for the content of this paper.