WHEN GOODNESS FORGETS SHAPE: INFORMATION LOSS IN FORWARD-FORWARD LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forward-Forward (FF) learning replaces end-to-end backpropagation with local goodness objectives, but local supervision is only useful if the goodness function preserves the information needed for prediction. We identify a failure mode in spatial representations: mean-squared goodness is invariant to permutations of hidden spatial positions, making some shape-relevant distinctions invisible to the local objective. Under spurious correlations, this invariance can suppress shape evidence while allowing simpler nuisance cues to dominate. We study spatial goodness through controlled adaptations that expose spatial structure to local supervision while preserving block-local training and avoiding cross-block gradients. Capacity-matched and coordinate-randomization controls isolate stable spatial indexing as the relevant source of improvement. On Colored MNIST, spatial goodness increases reversed-color accuracy from 10.83% to 85.02%. In a controlled adaptation of Adaptive Spatial Goodness Encoding (ASGE), spatial local predictions achieve 77.48% reversed-color accuracy, whereas global-average fusion of the same trained representation reduces performance to 10.12%, showing that spatial evidence can be learned locally yet discarded during prediction. Experiments on Waterbirds extend the analysis to natural images and improve worst-group performance. These results show that robust local learning depends not only on what representations encode, but also on what local objectives and prediction rules preserve.

1 INTRODUCTION

Forward-Forward (FF) learning trains neural networks through local goodness objectives instead of end-to-end backpropagation (BP) (Hinton, 2022). Each objective encourages a layer to distinguish positive data from negative data using its own activations. Local updates reduce dependence on a backward signal propagated through the entire network. This property motivates learning systems with constraints on activation storage, update coordination, or backward models. Subsequent work has extended FF to convolutional architectures and improved its scaling, contrastive training, and coordination across layers (Papachristodoulou et al., 2024; Krutsylo, 2025; Chen et al., 2025; Yao et al., 2026). As local learning scales, which information each local objective preserves becomes an increasingly consequential design choice.

Goodness determines which distinctions a local learner can use to separate data. This matters when the training distribution offers several predictive cues. A global color cue may identify a class as reliably as the object's shape during training. When the color association changes, a learner that depends on color can fail despite high in-distribution accuracy. Such shortcut reliance is well established in deep learning (Geirhos et al., 2020; Sagawa et al., 2020). In FF, it raises a specific question: can the reduction defining goodness obscure spatial evidence while keeping a shortcut visible? Answering this question requires examining how the objective measures a representation, alongside what that representation contains.

We study mean-squared activation goodness, a common reduction over channels and spatial positions. Its value records activation energy while discarding where that energy occurs. We formalize this observation through invariance to permutations of hidden spatial positions. Two activation tensors related by such a permutation receive identical goodness. When a class distinction is encoded only in their spatial arrangement, the score cannot distinguish them. A chromatic cue that changes activation

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.

energy can remain visible. The invariance applies to the goodness functional on hidden features; convolutional processing determines how input transformations affect those features. This distinction connects the mathematical property to a concrete question about shape learning under color shifts.

Prediction can discard spatial distinctions that the local heads have learned to use. A final classifier that globally averages block features may therefore discard spatial evidence retained by a local readout. We examine these two information bottlenecks using controlled adaptations of Adaptive Spatial Goodness Encoding (ASGE) and covariance-aware goodness (BiCovG) (Gong et al., 2026; Jiang et al., 2026). These objectives provide spatial supervision through different constructions. Comparing their local predictions with their fused predictions reveals whether useful evidence survives the path from local learning to the deployed decision. Figure 1 shows both information gates: the distinctions measured by local goodness and the evidence retained by prediction.

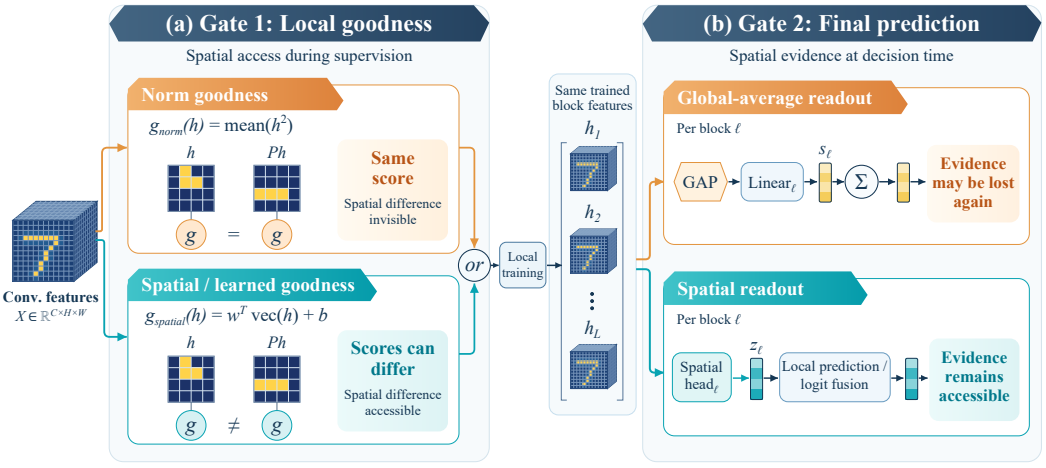

Figure 1: **Two-stage information loss of spatial evidence in local learning.** (a) Mean-squared norm goodness is invariant to spatial permutations, whereas an indexed spatial head can distinguish equal-energy arrangements. (b) The same spatially supervised block features can support different final decisions: global-average readout removes spatial indexing, while spatial readouts retain access to it. Panel (a) uses candidate-conditioned FF notation; panel (b) summarizes the readout choices examined with ASGE and BiCovG.

We investigate the first reduction by replacing norm pooling with learned spatial goodness heads while retaining block-local gradients. Each head assigns weights to spatially indexed activations, allowing equal-energy configurations to receive different scores. A controlled comparison of local and end-to-end training separates the effect of the training rule from that of the readout geometry. We then hold parameter count fixed with an invariant head and test the same spatial architecture with randomized coordinates. These controls distinguish the contribution of stable spatial indexing from additional head capacity. Under controlled color shifts, the resulting comparisons link the robustness improvement to the spatial distinctions available to local supervision.

We connect this analysis to a diagnostic of the training data. Sparse fuzzy rules identify interpretable concepts associated with class separation. A separate statistical screen compares color separation with a shape-carrying reference channel. Independent null environments set its threshold before evaluation. We study this diagnostic alongside targeted color diversification before retraining. The two experiments connect cue detection with changing training evidence, complementing the intervention on goodness geometry.

Our contributions are threefold:

- The invariance argument identifies spatial distinctions that mean-squared goodness cannot measure. Matched interventions connect this loss to shortcut reliance.
- Fusion analysis shows that prediction can discard spatial evidence already learned by local heads. Controlled color shifts isolate these mechanisms.

- Natural-image experiments trace their consequences through local adaptation and readout retraining. Together, these results show that robustness in FF learning depends not only on what representations encode, but also on what local objectives and prediction rules preserve.

2 Related Work

Goodness and local learning. FF trains layers through local goodness objectives (Hinton, 2022). Channel-wise competition (CwComp) extends convolutional FF without explicit negative construction (Papachristodoulou et al., 2024). Subsequent work improves scaling, self-contrastive training, and hierarchical coordination (Krutsylo, 2025; Chen et al., 2025; Yao et al., 2026). Goodness functions also differ in the information they measure. ASGE incorporates spatial energy (Gong et al., 2026), while BiCovG combines spatial scales, cross-channel covariance, and feature alignment (Jiang et al., 2026). Sparse goodness further studies selective activation measurement (Yuksel & Sawaf, 2026). These approaches modify how local objectives measure hidden representations, but the information preserved or discarded by the goodness reduction itself has received less attention. We focus on whether spatial evidence remains accessible to local supervision and subsequent prediction.

Spurious correlations and group robustness. Shortcut learning exploits correlations that change outside the training distribution (Geirhos et al., 2020). Overparameterization and simplicity bias help explain this preference (Sagawa et al., 2020; Yang et al., 2024), while group-robust evaluation exposes failures hidden by average accuracy (Koh et al., 2021). Prior approaches include retraining on initially misclassified examples (Liu et al., 2021), contrastive alignment without group labels (Zhang et al., 2022), and group-inference-based robust optimization (Han & Zou, 2024). Last-layer retraining further shows that representations can retain useful information even when predictions remain shortcut-driven (Kirichenko et al., 2023). Our setting complements this perspective by asking whether spatial evidence can be lost earlier, during local goodness computation, and again during final prediction.

Interpretable shortcut discovery and intervention. Interpretable proxies can identify evidence associated with shortcut behavior. Patch-based analyses expose classifier vulnerabilities (Rosenzweig et al., 2021), while prototype-based methods detect shortcut patches and support targeted retraining (Kuhn et al., 2025). Other work shows that suppressing one shortcut can strengthen another, highlighting the difficulty of identifying evidence that remains reliable across environments (Li et al., 2023). These methods primarily diagnose or mitigate shortcuts after they emerge in learned representations. We instead use interpretable diagnostics to connect shortcut reliance to the information made available by the local learning objective, together with controlled changes to the training evidence.

3 Information Loss in Local Learning

3.1 What the local objective can see

Goodness defines which distinctions receive supervision at each FF block. Consider an image x with class label $y \in \{1, \dots, C\}$. A candidate-conditioned classifier appends a candidate encoding c to the input. At block ℓ, the correct and incorrect candidates produce hidden activations h_ℓ^+ and h_ℓ^-. The block minimizes

$$\mathcal{L}_\ell = \mathrm{softplus}\big[-\big(g_\ell(h_\ell^+) - g_\ell(h_\ell^-)\big)\big]. \tag{1}$$

Each block is trained before the next block, and earlier blocks remain frozen. The choice of g_ℓ therefore shapes supervision without a gradient from the final decision.

A common goodness function reduces a feature map $h \in \mathbb{R}^{K\times H\times W}$ to its mean squared activation:

$$g_{\mathrm{norm}}(h) = \frac{1}{KHW} \sum_{k,u,v} h_{k,u,v}^2. \tag{2}$$

This score records activation energy but removes its spatial arrangement. Two feature maps can have the same energy while placing their responses at different locations. Their distinction is visible in the feature map and absent from the scalar objective.

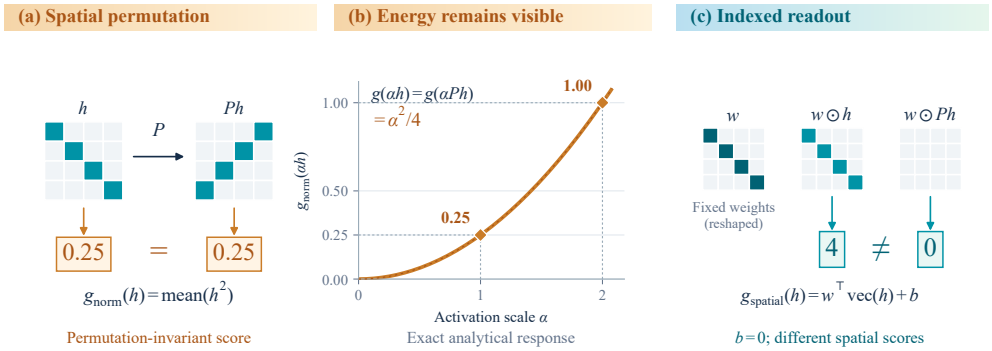

Figure 2: **Why norm goodness forgets arrangement.** An analytical example uses a single-channel 4×4 feature map with four unit activations; light cells are zero. (a) Spatial permutation preserves norm goodness. (b) The coincident analytical curves satisfy $g_{\text{norm}}(\alpha h) = g_{\text{norm}}(\alpha P h) = \alpha^2/4$, with marked values 0.25 at $\alpha = 1$ and 1 at $\alpha = 2$; they are not experimental fits. (c) A fixed diagonal weight pattern, shown as a reshaped vector, separates the maps through an indexed head. Elementwise products sum to 4 and 0 with $b = 0$. The head retains block-local gradients.

Proposition 1 (Spatial-permutation invariance)**.** *Let π be any permutation of the HW spatial positions, applied identically to every channel, and let $P_\pi h$ denote the resulting feature map. Then* $g_{\text{norm}}(Ph) = g_{\text{norm}}(h)$.

Proof. A spatial permutation reorders the squared entries in Equation 2. Their sum and the normalization remain unchanged.

The proposition applies to hidden features. If a class distinction is encoded only in their spatial arrangement, the goodness score cannot express it. A color cue that changes channel energy remains visible and may dominate during training. Reversing that cue at test time exposes the resulting shortcut reliance. Figure 2 gives a concrete hidden-feature example of this equivalence.

3.2 Preserving Spatial Distinctions

A spatial head gives local supervision access to feature locations. For candidate-conditioned FF, we replace norm goodness with

$$g_{\text{spatial}}(h) = w^\top \text{vec}(h) + b. \tag{3}$$

The head and its block are trained jointly with Equation 1. Earlier blocks stay detached. Position-dependent weights allow the score to distinguish arrangements that norm pooling treats as equal.

Stable spatial indexing is the mechanism tested by this intervention. We compare the spatial head with a parameter-matched MLP operating on per-channel energies. The MLP increases capacity while keeping spatial invariance. We also randomize coordinates independently at each evaluation of the same spatial head. This preserves the head architecture but removes a consistent mapping between positions and weights. Together, these controls separate stable spatial information from additional readout capacity.

3.3 Preserving Evidence Through Prediction

Local supervision and final prediction can use different summaries of the same representation. Let z_ℓ denote the class logits produced by a spatial local head. A fusion rule may combine these logits or instead classify globally averaged block features. The latter removes spatial indexing before making the final decision. Figure 1 therefore tracks both reductions: the local training score and the deployed readout.

We study this distinction using controlled adaptations of current goodness objectives. Channel-wise competition (CwComp) forms class logits from squared activations within class-specific channel groups (Papachristodoulou et al., 2024). Adaptive Spatial Goodness Encoding (ASGE) projects

spatial energy vectors into local class logits (Gong et al., 2026). Its fusion head instead reads global-average features from each detached block. BiCovG combines spatial energy descriptors across scales with cross-channel probes (Jiang et al., 2026). Its final predictor fuses layer logits. These readout choices let us test which parts of locally learned evidence survive prediction.

4 Concept Auditing and Targeted Cue Diversification

Spatial readouts change what a learner can distinguish; cue diversification changes which correlations it encounters. An interpretable audit connects these interventions through 15 channel-region means. A sparse fuzzy rule model identifies concepts associated with class separation. A separately calibrated statistic compares color separation against a blue-channel reference that preserves the same digit shape. Subtracting this reference separates added color evidence from separation already carried by shape. Appendix F defines the concepts, rule model, and reference-adjusted score.

Calibration fixes the decision threshold before evaluation on the target seeds. It uses held-out concept scores from grayscale and label-independent color, with 20 independent seeds per null environment. Targeted diversification tests the data intervention: randomize color, retain labels, and retrain the original FF objective. We evaluate the calibrated diagnostic and diversification separately to distinguish detecting cue dominance from changing training evidence.

5 Experiments

5.1 Study design

The experiments trace spatial evidence from local supervision to final prediction. Colored MNIST isolates shape and color under a controlled shift. Colored CIFAR-10 tests the same intervention on natural object images. Waterbirds and CelebA examine local adaptation and prediction with naturally occurring spurious attributes.

For Colored MNIST, digits 0–4 and 5–9 form two classes. Red or green agrees with the label with probability ρ during training. Reversed-color tests use $1 - \rho$, while random-color tests use 0.5. The primary geometry comparison uses 10,000 training images, 2,000 test images, and 20 epochs. Capacity controls and objective adaptations use 30,000/5,000 images and 15 epochs per block, with $\rho = 0.9$. Each comparison shares its data protocol across the reported learners.

All reported aggregate results use five seeds and show mean $\pm$ sample standard deviation. Paired differences compare matching seeds. Appendix J contains statistical tests and confidence intervals. Appendix A records architectures, optimizers, selection criteria, and hardware. The CwComp, ASGE, and BiCovG results use controlled objective adaptations in the common two-block setting.

5.2 Spatial access changes shortcut reliance

Figure 3 summarizes the main color-shift comparison, the correlation sweep, and the controls for stable spatial indexing. The choice of local reduction strongly affects reliance on color versus shape after the shift. At $\rho = 0.9$, norm FF reaches 10.83% reversed-color accuracy and falls near chance when color is randomized. Spatial FF reaches 85.02% and 90.79%, respectively (Table 1), yielding a paired reversed-color gain of 74.19 percentage points. Under random color, the assigned color cue carries no systematic label correlation, so the remaining performance indicates that the spatial model has learned predictive evidence beyond that cue.

A matched BP comparison separates this effect from the local training rule. BP with global-average pooling similarly relies on color, reaching 11.45% reversed-color accuracy, whereas its spatial counterpart reaches 79.28%. Thus, retaining spatial information substantially improves reversed-color performance under both FF and BP. The remaining difference between FF-Spatial and BP-Spatial is not statistically resolved by the present comparison.

Capacity and coordinate controls further isolate the role of stable spatial indexing. Table 2 compares invariant, randomized-coordinate, and stable-coordinate heads, each with 9,938 trainable parameters. Their reversed-color accuracies are 33.66%, 24.24%, and 90.64%, respectively. The stable-coordinate model outperforms both controls in every seed. Additional capacity therefore improves the invariant

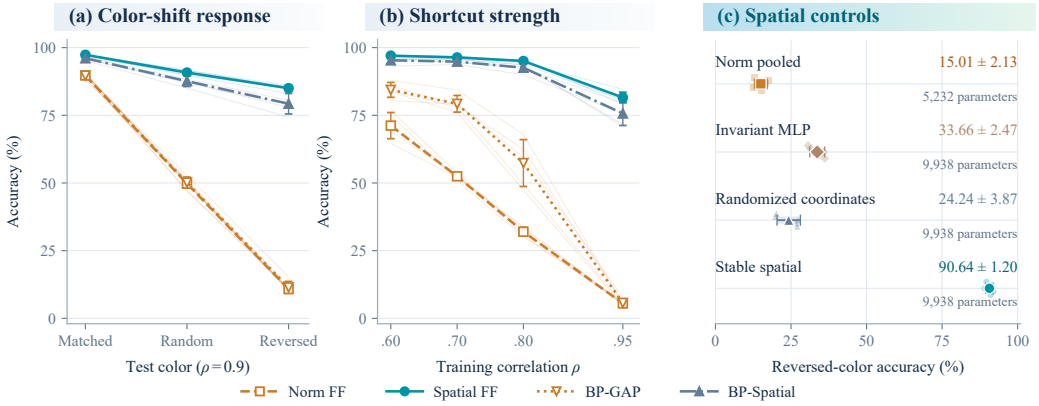

Figure 3: **Spatial indexing changes robustness under color shifts.** (a) Main comparison at $\rho = 0.9$. (b) Reversed-color accuracy across training correlations. (c) Capacity and coordinate controls at $\rho = 0.9$, with parameter counts shown. Panel (a) uses 10,000/2,000 training/test images and 20 epochs; panels (b) and (c) use 30,000/5,000 and 15 epochs. Local schedules apply per block. Large markers and bars show five-seed means and sample standard deviations. Faint traces connect matching seed indices; each correlation in (b) uses separately trained models. Faint seed points in (c) have small vertical display offsets to separate overlaps.

Table 1: Five-seed accuracy on Colored MNIST. Random color measures shape learning; reversed color measures shortcut robustness. The primary comparison is FF versus FF-Spatial at $\rho = 0.9$.

Learner	ρ	Random color	Reversed color
FF (norm)	0.5	74.40 ± 1.27	74.55 ± 1.48
FF-Spatial	0.5	96.19 ± 0.23	96.12 ± 0.16
BP-GAP	0.5	83.31 ± 0.70	83.16 ± 1.04
BP-Spatial	0.5	94.86 ± 0.52	94.83 ± 0.40
FF (norm)	0.9	49.75 ± 1.51	10.83 ± 0.51
FF-Spatial	0.9	90.79 ± 0.95	85.02 ± 1.11
BP-GAP	0.9	50.16 ± 2.04	11.45 ± 2.20
BP-Spatial	0.9	87.62 ± 1.98	79.28 ± 3.78

readout only partially, whereas the much larger gain appears when a consistent mapping between feature locations and head weights is preserved. Appendix B.4 provides a complementary post-training permutation-and-restoration test of this dependence. Appendix E further reports paired-color representation and prediction diagnostics.

Table 2: Capacity and coordinate controls at $\rho = 0.9$ (30,000/5,000 examples, 15 epochs, five seeds). Models with the invariant MLP or either spatial head have 9,938 trainable parameters, including 4,706 in their readouts.

Goodness head	Parameters	Random color	Reversed color
Norm pooled	5,232	52.96 ± 2.13	15.01 ± 2.13
Invariant MLP	9,938	63.44 ± 1.69	33.66 ± 2.47
Randomized spatial	9,938	58.05 ± 2.35	24.24 ± 3.87
Stable spatial	9,938	94.50 ± 0.63	90.64 ± 1.20

The same interaction holds across the tested color correlations. At $\rho \in \{0.6, 0.7, 0.8, 0.95\}$, spatial FF gains 25.83, 43.92, 63.06, and 76.04 paired points over norm FF. Figure 3b traces these changes, with paired gains in Appendix Figure A6. Colored CIFAR-10 also improves, from 9.99% to 20.31% reversed-color accuracy (Table A3). The smaller natural-object gain shows that preserving configuration changes learning without fully resolving the harder recognition task.

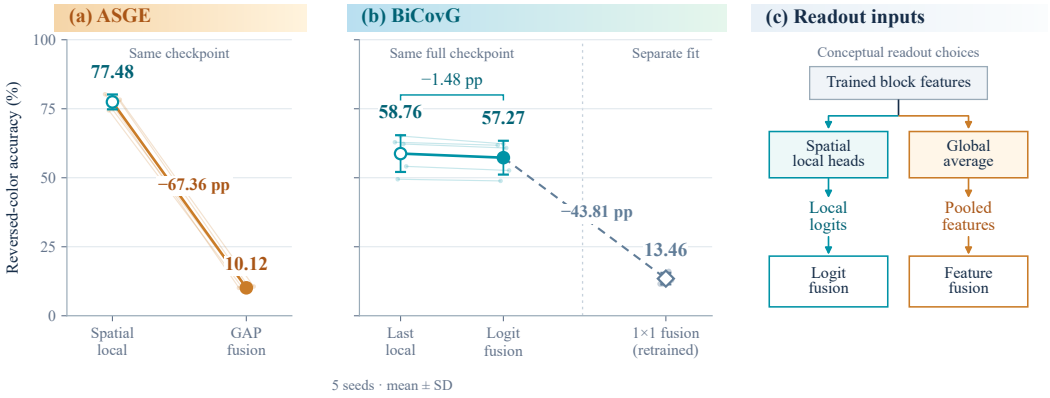

Figure 4: **Prediction can discard locally available spatial evidence.** (a) ASGE local and global-average fusion predictions from the same checkpoint. (b) BiCovG last-local and logit-fusion predictions from the same full checkpoint; the 1×1 single-scale model is trained separately. Faint lines pair seeds within a checkpoint; large markers and bars show five-seed means and sample standard deviations. Annotated changes are right minus left. (c) Logit fusion reads spatial-local decisions; feature fusion reads global averages. ASGE and BiCovG have separate checkpoints. Results use 30,000/5,000 images and $\rho = 0.9$.

5.3 Final prediction can discard locally learned evidence

Figure 4 compares local and fused predictions obtained from the same trained blocks. Spatial supervision and deployed prediction can behave very differently under the same distribution shift. In the ASGE adaptation, the final spatial-local logits achieve 77.48% reversed-color accuracy, as shown in Figure 4a. Global-average fusion of the same trained representation reaches only 10.12%. The paired difference of 67.36 percentage points is positive in every seed. Thus, useful evidence remains accessible to the spatial local head even though the classifier operating on globally averaged features loses most of this robustness. Appendix D.3 further tests this distinction with separately fitted readouts on frozen ASGE blocks under two probe-training regimes. Spatial resolution also affects the local decision itself. An ASGE grid pair of $4 \times 4/2 \times 2$ retains 70.56% reversed-color accuracy in its last local head. The $2 \times 2/1 \times 1$ pair falls to 10.54%. These results connect both local supervision and feature fusion to the spatial distinctions they retain.

BiCovG provides a complementary test because it fuses local logits. Its fused and final-local accuracies are 57.27% and 58.76%, a paired gap of 1.48 points. Restricting both blocks to a single 1×1 scale reduces fused accuracy by 43.81 paired points. Removing cross-channel probes or feature alignment produces no statistically resolved change in this setting. The contrast with ASGE links prediction behavior to how spatial evidence is summarized across layers.

Stronger local competition also improves over scalar norm goodness. The CwComp adaptation reaches 74.55% reversed-color accuracy, compared with 14.91% for pooled candidate-conditioned FF. Spatial FF reaches 90.80% under this protocol. Together, these objective and readout comparisons locate distinct sources of information loss in local learners. Appendix Table A4 gives the full objective comparison and training times.

5.4 Cue detection supports targeted diversification

The calibrated audit separates cue dominance from ordinary class separation in the controlled protocol. It triggers in every tested seed at $\rho \geq 0.6$ and in none at $\rho = 0.5$ or in grayscale controls. The reference channel matters because digit shape can itself separate the classes. Subtracting its effect keeps that separation from being treated as added color evidence.

The data intervention tests whether removing the named correlation changes the learner's decision. With deterministic training color, norm FF reaches 100% matched-color accuracy and 0% reversed-color accuracy. Targeted color diversification raises reversed-color accuracy to $84.08 \pm 1.21\%$. Always randomizing color reaches $83.50 \pm 1.54\%$. The two interventions have similar observed means;

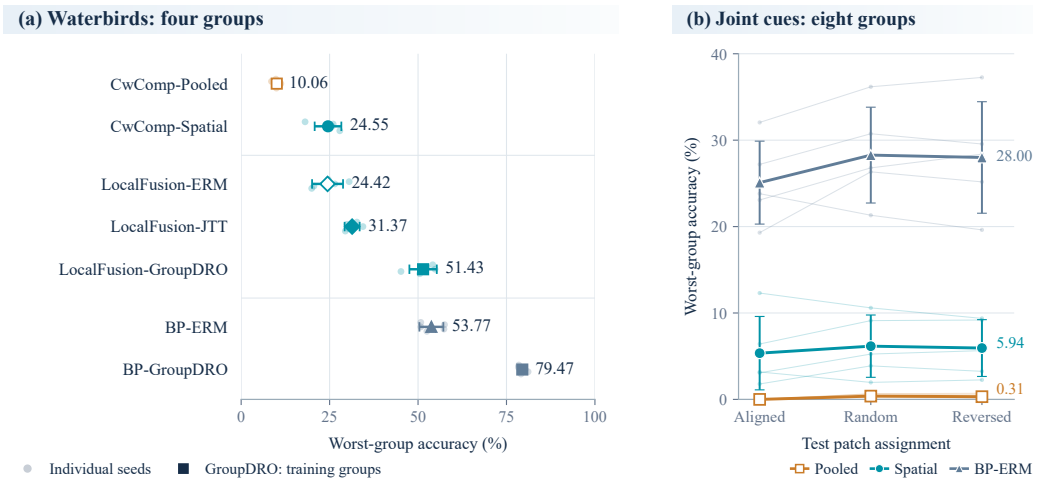

Figure 5: **Spatial evidence on natural images with correlated cues.** (a) Waterbirds worst-group accuracy across the four label–background groups. (b) Worst-group accuracy across eight label–background–patch groups under aligned, random, and reversed patch assignments. Markers and error bars denote five-seed means and sample standard deviations, while faint points and traces show individual seeds. Pooled and Spatial denote CwComp-Pooled and CwComp-Spatial. Filled squares indicate GroupDRO with training-group labels. Local checkpoints use validation worst-group accuracy; JTT fusion uses no group labels during optimization or selection.

calibration separately prevents triggers in the tested null environments. Appendix I distinguishes the calibrated screen from the trigger used in the repair experiment.

5.5 Spatial evidence in natural-image learning

Figure 5 compares the spatial readout effect on Waterbirds with its behavior under a second correlated cue. Waterbirds tests whether the spatial interaction survives updates to a pretrained visual representation. Bird labels correlate with land or water backgrounds in this dataset (Sagawa et al., 2020). Worst-group accuracy takes the minimum across its four label-background groups. All models start from the same ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) and use the official data splits. The local learners greedily update four residual stages, without gradients crossing stage boundaries. CwComp-Spatial replaces class-channel means with class-specific heads over a 4×4 grid of squared activations. Local checkpoints are selected by validation worst-group accuracy. This setup compares the two reductions during local adaptation from image pixels.

Table 3: Waterbirds adaptation from an ImageNet-pretrained ResNet-18 over five seeds. LocalFusion combines four detached spatial-local logit pairs. GroupDRO uses training group labels, and local-stage selection uses validation group labels. Values are percentages.

Method	Average acc.	Worst-group acc.
CwComp-Pooled	57.46 ± 0.69	10.06 ± 1.01
CwComp-Spatial	72.61 ± 1.35	24.55 ± 3.77
LocalFusion-ERM	67.39 ± 5.49	24.42 ± 4.38
LocalFusion-JTT	77.27 ± 1.92	31.37 ± 2.12
LocalFusion-GroupDRO	70.89 ± 2.30	51.43 ± 3.87
BP-ERM	81.75 ± 1.30	53.77 ± 3.39
BP-GroupDRO	85.98 ± 0.73	79.47 ± 1.19

Table 3 shows that spatial goodness raises worst-group accuracy from 10.06% to 24.55%. The gain is positive in every seed. The same pattern also appears when the representation is frozen: spatial FF improves worst-group accuracy from 51.71% to 58.50%. Appendix Table A9 reports the complete

frozen-representation comparison. Together, these settings show that the spatial-readout effect is present both in fixed features and in locally adapted features.

Frozen local logits further reveal how the final objective uses evidence distributed across stages. A linear ERM fusion head reaches 24.42% worst-group accuracy, close to the final spatial-local head. JTT retrains this fusion head with greater weight on examples misclassified by the initial predictor and reaches 31.37%, without using group labels during fusion optimization or model selection. Group-supervised fusion reaches 51.43%, compared with 53.77% for global ERM and 79.47% for global GroupDRO. These results indicate that useful evidence remains distributed across local stages, although exploiting it depends strongly on the final training objective.

Removing group labels from local-stage selection provides a stricter test of the fully groupless pipeline. Using average-accuracy selection throughout, JTT reaches 28.50% worst-group accuracy. Its 6.32-point improvement over the corresponding local head is not resolved statistically by the present comparison. On CelebA, the same fully groupless procedure reaches 42.78%, compared with 44.33% for the spatial-local head. Thus, the fusion improvement observed on Waterbirds does not consistently transfer to this second task. Appendix H reports the selection rules and complete results.

5.6 Distinguishing information retention from cue selection

Multiple shortcuts test a different requirement from retaining spatial configuration. We add a label-correlated red or green corner patch to Waterbirds, while leaving the natural background intact. The minimum accuracy across eight label-background-patch groups measures reliance on either cue. Spatial goodness improves over pooled goodness in every seed and environment, but the spatial model still reaches only about 5–6% joint worst-group accuracy, while the pooled model remains near zero. Figure 5b shows this behavior across the three patch environments, and Appendix Table A13 reports the corresponding numerical results. Global ERM reaches 25.09–28.27% across these environments. Spatial access therefore helps the local learner without determining which cue remains reliable after a shift.

In the single-seed deterministic-color test, every learner reaches 100% matched-color accuracy and 0% reversed-color accuracy. Training supplies no counterexamples to color, and spatial access alone does not change the learned decision. Appendix I reports the CIFAR-10 location-cue comparison, which is also unresolved. Its input-level perturbation need not produce a spatial permutation of hidden activations.

6 Conclusion

Goodness determines which information receives local supervision, and prediction fusion determines which information reaches the final decision. Mean-squared goodness merges spatial arrangements with equal activation energy. Controlled interventions connect this invariance to color-shortcut reliance and show the value of stable spatial indexing. The fusion experiments reveal a second loss: useful spatial-local decisions can disappear after global feature pooling. Concept auditing and targeted diversification provide a complementary way to change the evidence available for learning. Across these results, preserving spatial information through both training and prediction emerges as a concrete design principle for FF learning.

Reproducibility Statement

Implementation details, model architectures, optimization settings, data construction, checkpoint selection, and statistical analyses are provided in Appendix A and the corresponding experiment-specific appendices. The supplementary package provides an anonymous standalone implementation of the rule-based goodness feasibility study, including seeded experiments, result summaries, extracted rules, and run metadata. Complete protocols for the image-based experiments are documented in the appendices.

AI USE STATEMENT

Generative AI tools assisted with data analysis, manuscript editing, and figure preparation. The authors independently verified all AI-assisted mathematical claims, experimental results, citations, and conclusions and take full responsibility for the final work.

REFERENCES

Xing Chen, Dongshu Liu, Jérémie Laydevant, and Julie Grollier. Self-contrastive forward-forward algorithm. *Nature Communications*, 16(1):5978, 2025. doi: 10.1038/s41467-025-61037-0.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Robert Geirhos, Jorn-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. doi: 10.1038/s42256-020-00257-z.

Qingchun Gong, Robert Bogdan Staszewski, and Kai Xu. Adaptive spatial goodness encoding: Scaling the forward-forward algorithm for convolutional neural networks. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2446–2450, 2026. doi: 10.1109/ICASSP55912.2026.11464145. URL https://doi.org/10.1109/ICASSP55912.2026.11464145.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 17480–17504, 2024. URL https://proceedings.mlr.press/v235/han24g.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022. URL https://arxiv.org/abs/2212.13345.

Xiaoyi Jiang, Bashir M. Al-Hashimi, and Kai Xu. Covariance-aware goodness for scalable forward-forward learning. *arXiv preprint arXiv:2605.04346*, 2026. URL https://arxiv.org/abs/2605.04346.

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*, 2023. URL https://mlanthology.org/iclr/2023/kirichenko2023iclr-last/.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Andrii Krutsylo. Scalable forward-forward algorithm. *arXiv preprint arXiv:2501.03176*, 2025. URL https://arxiv.org/abs/2501.03176.

Lukas Kuhn, Sari Sadiya, Jörg Schlötterer, Florian Buettner, Christin Seifert, and Gemma Roig. Efficient unsupervised shortcut learning detection and mitigation in transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2217–2226, 2025.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_A_Whac-a-Mole_Dilemma_Shortcuts_Come_in_Multiples_Where_Mitigating_One_CVPR_2023_paper.html.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6781–6792, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Andreas Papachristodoulou, Christos Kyrkou, Stelios Timotheou, and Theocharis Theocharides. Convolutional channel-wise competitive learning for the forward-forward algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14536–14544, 2024. URL https://openreview.net/forum?id=ZAW5wYSLhS.

Julia Rosenzweig, Joachim Sicking, Sebastian Houben, Michael Mock, and Maram Akila. Patch shortcuts: Interpretable proxy models efficiently find black-box vulnerabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 56–65, 2021.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2953–2961, 2024. URL https://proceedings.mlr.press/v238/yang24c.html.

Jie-En Yao, Hong-En Chen, and C.-C. Jay Kuo. HCL-FF: Hierarchical and contrastive learning for forward-forward algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27429–27438, 2026.

Kamer Ali Yuksel and Hassan Sawaf. Sparse goodness: How selective measurement transforms forward-forward learning. *arXiv preprint arXiv:2604.13081*, 2026. URL https://arxiv.org/abs/2604.13081.

Michael Zhang, Nimit S. Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516, 2022. URL https://proceedings.mlr.press/v162/zhang22z.html.

A Implementation and Evaluation Protocols

This appendix provides the implementation and evaluation details required to reproduce the controlled and natural-image experiments. The protocols are organized around the two information bottlenecks studied in the main text: information removed by local goodness computation and information removed by downstream prediction. Controlled image experiments manipulate cue availability while keeping the underlying task labels fixed, whereas the natural-image experiments compare local readouts, block-wise adaptation, and prediction fusion under explicitly specified checkpoint-selection rules. This section defines settings shared across experiments; subsequent appendices describe mechanism-specific controls, ablations, and full results.

A.1 Data construction and evaluation

Colored MNIST. We construct Colored MNIST from the official MNIST training and test splits (LeCun et al., 1998), using binary labels: digits 0–4 form class 0 and digits 5–9 form class 1. For each seed, images are sampled without replacement from the corresponding split and pixel intensities are rescaled to $[0, 1]$ by division by 255. Within each experimental comparison, all learners use the same sampled images and labels, eliminating variation due to example selection.

For a grayscale image s with binary label y, color is assigned independently with agreement probability ρ. Let $a \sim \mathrm{Bernoulli}(\rho)$ and define

$$c = \begin{cases} y, & a = 1, \\ 1 - y, & a = 0. \end{cases}$$

Channel c receives s, the other red/green channel is set to zero, and the blue channel receives $0.05s$. The blue channel therefore preserves digit-shape information while remaining independent of the red/green color assignment. For the clean-image control, s is replicated across all three channels.

Evaluation environments. We evaluate three color environments while preserving the digit-label task. The matched-color environment uses the training agreement probability ρ, the random-color environment uses $\rho = 0.5$, and the reversed-color environment uses $1 - \rho$. Color assignments are generated from independent random draws. Reversed-color accuracy measures robustness when the training correlation changes direction, whereas random-color accuracy measures performance when the assigned color carries no systematic label information.

Sample sizes and training schedules. The primary geometry experiment uses 10,000 training images, 2,000 test images, and 20 training epochs. Capacity-matched controls and the modern-objective comparisons use 30,000 training images, 5,000 test images, and 15 epochs. For FF models, the stated number of epochs is applied independently to each locally trained block; for BP baselines, the same number denotes end-to-end training epochs. Unless otherwise stated, all aggregate experiments use five seeds, indexed 0 through 4.

Table A1: Controlled-task protocols. Local epochs apply to each block. All rows use five seeds and batch size 256.

Experiment	Train / test	Widths	Epochs	Learning rate
MNIST geometry	10,000 / 2,000	16, 32	20	0.002
MNIST capacity	30,000 / 5,000	16, 32	15	0.002
MNIST correlation sweep	30,000 / 5,000	16, 32	15	0.002
CwComp adaptation	30,000 / 5,000	16, 32	15	0.002
ASGE adaptation	30,000 / 5,000	16, 32	15	0.0002
BiCovG adaptation	30,000 / 5,000	16, 32	15	0.0002
Colored CIFAR-10	10,000 / 2,000	32, 64	10	0.0005

A.2 Candidate-conditioned models

Candidate-conditioned FF appends a constant label plane to the three input channels. The positive plane contains the binary target; the negative plane contains its complement. Each block uses a 3×3

convolution with unit padding, ReLU, and 2×2 max pooling. The two MNIST blocks produce $16 \times 14 \times 14$ and $32 \times 7 \times 7$ activations. The second block receives detached features from the frozen first block.

Training compares positive and negative goodness with Equation 1. At prediction time, the network evaluates both candidate planes and sums their goodness across blocks. The candidate with the larger sum determines the prediction. Norm FF uses squared activation means; spatial FF uses one learned linear scalar head per block. Both use AdamW with learning rate 0.002 and batch size 256. The implementation retains AdamW's default weight decay of 0.01.

The BP references share the two convolutional widths and pooling operations. BP-GAP averages the final activation map before its two-class linear head. BP-Spatial retains a 4×4 grid through adaptive average pooling. Their training loss is cross-entropy, and both use the same AdamW schedule as the primary FF comparison. The total parameter counts are 5,232 for norm FF and 9,938 for spatial FF. BP-GAP and BP-Spatial contain 5,154 and 6,114 parameters, respectively.

A.3 Computing Environment

Experiments were run primarily on NVIDIA RTX A6000 GPUs. The software environment used Python 3.11, PyTorch 2.5.1, torchvision 0.20.1, and CUDA 12.1. Statistical analyses used NumPy 1.26.4, pandas 2.3.3, and SciPy 1.16.3; Waterbirds experiments used WILDS 2.0.0.

Random seeds determine dataset sampling, cue assignment, and model initialization. For the natural-image experiments, NumPy and PyTorch random seeds are fixed, but CUDA deterministic algorithms are not enforced. Accordingly, exact bitwise reproduction across hardware or software environments is not guaranteed; the reported multi-seed summaries are the intended unit of reproducibility. Training times reported in Table A4 are wall-clock measurements obtained on the shared experimental machine.

B Geometry and Capacity Controls

B.1 What the permutation result identifies

Norm goodness assigns the same score to activation maps that differ only by a spatial reordering. For each channel, a spatial permutation changes coordinate identity while preserving the multiset of activation values; summing squared activations therefore yields the same scalar. This equivalence applies to hidden activation maps, which are the direct inputs to the goodness function.

A convolutional representation can change more than coordinates when the input moves. Padding, receptive fields, and nonlinearities can change its activation values. The proposition therefore identifies information removed by the reduction after feature formation. The location-cue experiment in Appendix I tests this distinction at the input level.

Spatial goodness makes coordinate identity available to the local objective. Writing $h_{k,p}$ for channel k at position p, its linear score is

$$g_{\text{spatial}}(h) = \sum_{k,p} w_{k,p} h_{k,p} + b. \tag{4}$$

A spatial permutation can change this sum when position weights differ. The local block can then receive different learning signals for distinct arrangements of equal-energy activations. This capability motivates the two controls below.

B.2 A parameter-matched invariant head

The invariant control increases readout capacity without exposing spatial coordinates. It first computes the mean squared activation of each channel,

$$e_k(h) = \frac{1}{HW} \sum_{u,v} h_{k,u,v}^2. \tag{5}$$

A two-layer MLP maps $e(h)$ to scalar goodness. Its hidden widths are 131 for block one and 69 for block two. These heads contain 2,359 and 2,347 parameters. Their combined 4,706 parameters exactly match the two spatial linear heads.

The full invariant model therefore contains 9,938 parameters, matching both spatial controls in Table 2. At reversed color, it reaches 33.66%, compared with 15.01% for norm goodness. Stable spatial goodness reaches 90.64% under the same protocol. The invariant MLP improves the use of channel energy, while the remaining gap isolates the value of spatial configuration.

B.3 A COORDINATE-RANDOMIZED SPATIAL HEAD

The coordinate control keeps the spatial architecture and changes the consistency of its inputs. Each goodness evaluation independently permutes spatial positions for every example. The permutation is shared across channels, preserving the channel vector at each moved location. The convolutional output passed to the next block remains unshuffled.

Training uses new permutations for positive and negative goodness evaluations. At inference, both candidate labels receive matched permutations through a reset random-generator state. This matching prevents candidate-specific random draws from favoring one label. The model still has 9,938 parameters, while its head receives unstable coordinates across evaluations.

Randomized coordinates yield 24.24% reversed-color accuracy, compared with 90.64% using stable coordinates. The paired improvement is positive for all five seeds. Taken together, the controls separate spatial indexing from parameter count and readout flexibility. Statistical contrasts appear in Appendix J.

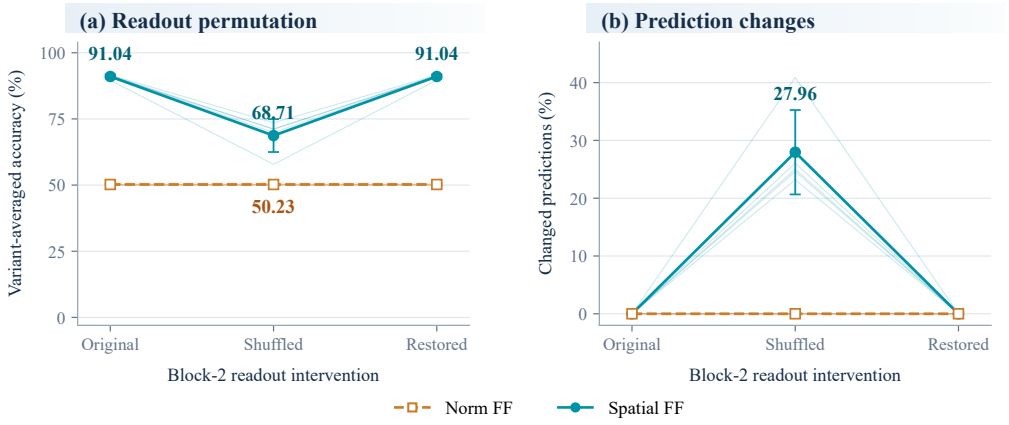

Figure A1: **Spatial-index intervention at the goodness readout.** Primary-protocol reruns evaluate red and green variants of 2,000 test images per seed at training color agreement $\rho = 0.9$. Only the second-block goodness readout is permuted; the first-block contribution and propagated features remain fixed. Restoration co-permutes spatial weights and feature coordinates without retraining. (a) Accuracy averaged over both variants. (b) Prediction-change rate relative to the original condition. Twenty permutation draws are averaged within each seed. Markers and error bars show five-seed means and sample standard deviations; faint traces show individual seed results.

B.4 COORDINATE CORRESPONDENCE IN A TRAINED GOODNESS HEAD

We perform a complementary post-training intervention to test whether prediction depends on the learned correspondence between feature coordinates and spatial-head weights. Primary-protocol reruns at $\rho = 0.9$ evaluate both red and green variants of each of 2,000 test images per seed. Only the second block's goodness readout is permuted. The first block's score contribution and the features propagated between blocks remain fixed. For the spatial head, jointly reordering its weights and feature coordinates restores the original correspondence. This intervention separates test-time readout dependence from the preceding control, which randomizes coordinates throughout training.

Figure A1 shows that spatial FF's variant-averaged accuracy decreases from 91.04% to 68.71% under the permutation and returns to 91.04% after coordinate restoration. The shuffled readout

changes 27.96% of predictions; restoration returns them to their original values. Norm FF remains at 50.23%, consistent with permutation-invariant goodness. The 22.33-point accuracy decrease and the prediction-change rate measure different consequences of the same intervention. These paired-color accuracies average two variants per base image; the reversed-color accuracies of 10.83% and 85.02% in Table 1 use a different color-assignment distribution.

C Correlation Strength and Natural-Image Extension

C.1 A common protocol across shortcut strengths

The correlation sweep measures how the geometry interaction changes with cue reliability. It uses 30,000 training images, 5,000 test images, 15 epochs, and the same four learners throughout. Table A2 gives the complete reversed-color results for the four matched correlations. This common schedule separates the sweep from the smaller primary experiment in Table 1.

Table A2: Reversed-color accuracy across the matched correlation sweep. Values are five-seed percentages with sample standard deviations.

Learner	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.95$
Norm FF	71.19 ± 4.83	52.47 ± 1.09	32.00 ± 1.63	5.52 ± 0.33
Spatial FF	97.02 ± 0.31	96.38 ± 0.41	95.06 ± 0.39	81.56 ± 1.97
BP-GAP	84.40 ± 2.75	79.25 ± 3.07	57.37 ± 8.66	5.72 ± 1.10
BP-Spatial	95.30 ± 1.22	94.82 ± 0.76	92.56 ± 1.54	75.67 ± 4.43

Spatial readouts maintain higher reversed-color accuracy as the training cue becomes more reliable. The spatial-minus-norm FF gains are 25.83, 43.92, 63.06, and 76.04 percentage points. Spatial BP follows the same pattern. At $\rho = 0.95$, both pooled learners closely follow color, while spatial learners still classify most reversed-color images correctly. The shared pattern connects the result to spatial reduction across both training rules.

The deterministic-color stress test removes all training counterexamples to the cue. In this separate seed-0 run, all four learners reach 100% matched-color accuracy and 0% reversed-color accuracy. It is excluded from five-seed inference. Spatial sensitivity supplies access to configuration, but counterexamples still influence which available cue training selects.

C.2 Colored CIFAR-10

We extend the controlled color-shift experiment to CIFAR-10, where the underlying images contain substantially richer natural-image structure than MNIST (Krizhevsky & Hinton, 2009). The binary task separates vehicles from animals: airplanes, automobiles, ships, and trucks form the vehicle class, and the remaining six CIFAR-10 categories form the animal class. Each split contains equal numbers of examples from the two binary classes, sampled without replacement.

RGB images are first converted to luminance using coefficients $(0.299, 0.587, 0.114)$. The assigned red or green channel receives the full luminance image, while each remaining channel receives 0.05 times that image. During training, color agrees with the binary label with probability $\rho = 0.9$. Matched-, random-, and reversed-color evaluations use agreement probabilities 0.9, 0.5, and 0.1, respectively.

Both learners use candidate-label planes and two convolutional blocks with 32 and 64 channels. The corresponding spatial heads operate on 16×16 and 8×8 activation maps. Models train on 10,000 examples and are evaluated on 2,000 test examples. Each block trains for 10 epochs with AdamW, learning rate 5×10^{-4}, weight decay 10^{-3}, and batch size 256. All optimization and data settings are shared between the pooled and spatial variants.

As shown in Table A3, Spatial FF improves reversed-color accuracy from 9.99% to 20.31%, a paired gain of 10.32 percentage points that is positive in every seed. Random-color accuracy increases from 49.82% to 55.54%. In contrast, matched-color accuracy differs by less than one percentage point, with both models exceeding 89%. The distribution shift therefore reveals a difference that is largely

Table A3: Colored CIFAR-10 at training correlation 0.9. Values are five-seed percentages with sample standard deviations.

Learner	Matched color	Random color	Reversed color
FF-Pooled	89.99 ± 0.69	49.82 ± 0.93	9.99 ± 0.54
FF-Spatial	90.69 ± 0.89	55.54 ± 0.90	20.31 ± 2.28

hidden by matched-color performance. The smaller gain relative to Colored MNIST indicates that preserving spatial indexing remains beneficial in this more complex image setting, while substantial residual dependence on color remains.

D Spatial Goodness Objectives and Prediction Fusion

The modern-objective comparison places three local learning designs in a shared two-block setting. CwComp and ASGE follow mechanisms in their released implementations. BiCovG follows its published equations. The resulting models use 16 and 32 channels on the controlled MNIST task. Their names refer to these objective adaptations throughout the results.

Table A4: Controlled objective adaptations on Colored MNIST at $\rho = 0.9$. All rows use 30,000/5,000 images and five seeds. Local blocks train for 15 epochs. Training times are measured on the same machine; local and fused ASGE/BiCovG predictions share their trained blocks.

Learner	IID	Random color	Reversed color	Train (s)
FF-Pooled	90.26 ± 0.55	52.89 ± 2.13	14.91 ± 2.15	36.1 ± 0.3
CwComp-adapted	95.45 ± 0.47	85.01 ± 1.99	74.55 ± 3.35	34.9 ± 1.4
ASGE fusion	89.75 ± 0.71	50.29 ± 1.00	10.12 ± 0.29	22.3 ± 1.0
ASGE spatial-local	96.05 ± 0.27	87.12 ± 1.42	77.48 ± 2.70	shared
BiCovG fusion	93.92 ± 0.34	75.73 ± 3.69	57.27 ± 6.13	29.9 ± 1.2
BiCovG last-local	93.94 ± 0.37	76.40 ± 3.98	58.76 ± 6.66	shared
FF-Spatial	98.07 ± 0.19	94.61 ± 0.58	90.80 ± 0.91	38.8 ± 1.1

D.1 Channel-wise competitive learning

CwComp assigns half of each block's channels to each class. A block applies convolution, ReLU, 2×2 max pooling, and BatchNorm. Its class logit is the mean squared activation within that class's channel group. Cross-entropy trains each block locally, and prediction uses the final block's two class logits.

The adaptation uses Adam with learning rate 0.002, following the released implementation's optimizer choice. Each block trains for 15 epochs with batch size 256. Candidate label planes and explicit negative examples are absent. CwComp therefore tests class competition while retaining spatially pooled class-channel goodness.

D.2 ASGE spatial energy and fusion

ASGE computes squared activation averages on a fixed spatial grid. Let Pool_q denote adaptive average pooling to a $q \times q$ grid. For block ℓ, its local logits are

$$z_\ell(h) = R_\ell^\top \operatorname{vec}\big(\mathrm{Pool}_{q_\ell}(h^2)\big) + b_\ell. \tag{6}$$

The projection and bias are fixed Gaussian draws scaled by $1/\sqrt{C}$, where $C = 2$. Local cross-entropy updates the convolution. The projection itself remains fixed.

Propagation uses average pooling followed by per-example RMS normalization and detachment. The next block therefore receives normalized features without a gradient path to the earlier block. Separate

linear classifiers read each block's detached global-average features. Their summed predictions form the fusion output.

Both local blocks and the fusion classifiers update within each minibatch for 15 epochs. They use separate AdamW optimizers with learning rate 0.0002 and weight decay 0.001. The main grid pair is 7×7 in block one and 4×4 in block two. All configurations contain 5,188 trainable parameters because the spatial projections are fixed.

The grid sweep localizes the loss to the information retained for each decision. Final local grids of 4×4 and 2×2 preserve high reversed-color accuracy. A 1×1 final grid sharply reduces it. Global-average fusion remains near 10% across all four configurations. Thus a useful spatial local prediction can coexist with a color-driven fusion prediction in the same trained model.

Table A5: ASGE grid sensitivity under the shared five-seed protocol. Local prediction uses the final block's fixed spatial logits. Values are percentages with sample standard deviations.

Block grids	Prediction	Matched color	Random color	Reversed color
$7 \times 7, 4 \times 4$	Local	96.05 ± 0.27	87.12 ± 1.42	77.48 ± 2.70
	Fusion	89.75 ± 0.71	50.29 ± 1.00	10.12 ± 0.29
$4 \times 4, 2 \times 2$	Local	95.42 ± 0.34	83.09 ± 1.50	70.56 ± 2.35
	Fusion	89.75 ± 0.71	50.30 ± 1.01	10.12 ± 0.29
$2 \times 2, 1 \times 1$	Local	89.82 ± 0.63	50.53 ± 1.26	10.54 ± 0.51
	Fusion	89.75 ± 0.71	50.30 ± 1.00	10.16 ± 0.29
$1 \times 1, 1 \times 1$	Local	90.11 ± 0.62	52.07 ± 2.41	13.40 ± 2.96
	Fusion	89.74 ± 0.71	50.30 ± 1.04	10.13 ± 0.29

D.3 Diagnostic readouts on frozen ASGE states

The frozen-state diagnostic holds the learned representation fixed while varying the readout and the color distribution used to fit it. Within each seed, four diagnostic heads are trained separately on the same frozen ASGE block states: GAP-linear, GAP-MLP, local-logits linear, and spatial-linear. The two conditions in Figure A2 differ only in the probe-training color distribution: one retains the training color–label correlation, whereas the other makes color independent of the label. The frozen ASGE backbone is unchanged between the two conditions. These diagnostic probes are fitted separately from the native ASGE local and fusion heads reported in Table A5.

Under correlated-color probe training, reversed-color accuracy ranges from 15.33% for GAP-linear to 90.31% for spatial-linear. When probe-training color is made label-independent, the corresponding accuracies rise to 71.11% and 95.45%. GAP-MLP similarly increases from 58.92% to 86.65%. Its 2,603 trainable parameters closely match the 2,594 parameters of the spatial-linear head, providing a comparison at similar readout capacity.

These results show that the frozen pooled representation retains predictive information that is not fully used by a color-correlated linear readout. At the same time, explicit spatial access yields the strongest reversed-color performance under both probe-training regimes. The diagnostic therefore distinguishes two contributors to prediction behavior in this setting: the information exposed by the readout and the decision rule learned from the probe-training distribution.

D.4 BiCovG components and scale ablations

BiCovG combines descriptors computed at multiple spatial scales before local classification. At each scale, squared activations are pooled independently within each channel. A learned 1×1 convolution additionally produces cross-channel probes whose squared responses are pooled over the same grids. With reduction ratio 8, this yields two probes in block 1 and four in block 2. A learned linear classifier operates on the concatenated multiscale descriptors.

Block 1 uses 2×2 and 4×4 grids, whereas block 2 uses 1×1 and 2×2 grids. Propagation applies RMS pooling, RMS normalization, and detachment. The feature-alignment module adds a

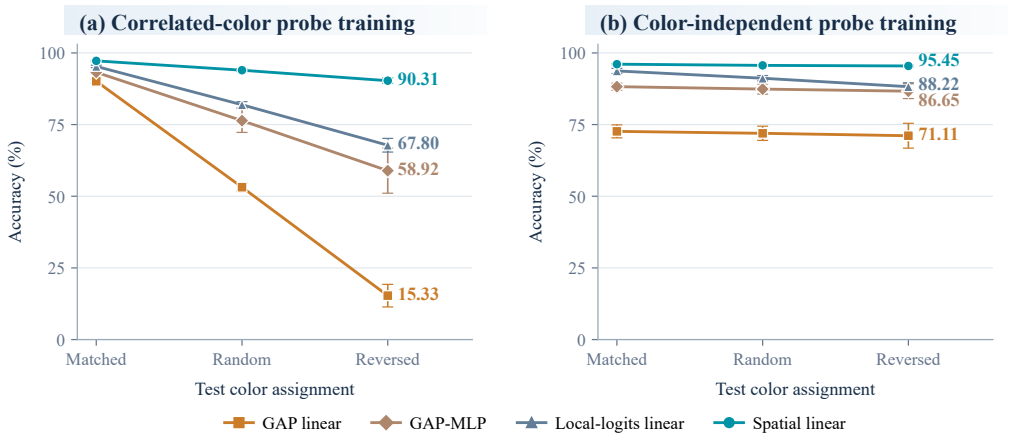

Figure A2: **Frozen-readout controls under two color-training regimes.** Four diagnostic heads are fitted separately to the same frozen ASGE states. (a) Probe training retains correlated color. (b) Probe training makes color independent of the label. Each head is evaluated under matched, random, and reversed test-color assignments. Points and error bars show five-seed means and sample standard deviations. GAP-MLP and spatial-linear use near-matched parameter budgets of 2,603 and 2,594. Readout fitting changes between panels; the backbone stays fixed within each seed. These probes are separate from the native ASGE local and fusion heads.

channel-wise offset derived from globally averaged features through a 512-unit hidden layer and a zero-initialized output projection.

Local components are trained for 15 epochs with AdamW, learning rate 2×10^{-4}, and weight decay 10^{-3}. The feature-alignment module is optimized jointly with block 2. After local training, the detached logits from the two blocks are combined using a softmax over two learned scalar fusion weights. These weights are optimized with Adam for 500 epochs at learning rate 0.01. Thus, unlike feature-level fusion, the BiCovG readout combines class evidence that has already been formed locally.

Table A6: BiCovG component ablations. Prediction columns report reversed-color accuracy over five seeds. Parameter counts describe each implemented configuration.

Configuration	Parameters	Final local	Fusion
Full mapping	23,246	58.76 ± 6.66	57.27 ± 6.13
Without cross-channel probes	23,126	53.04 ± 3.08	51.95 ± 3.14
Single 1×1 scale	22,274	13.68 ± 2.23	13.46 ± 2.08
Without feature alignment	6,334	59.44 ± 6.78	57.98 ± 6.49

Table A6 shows that spatial scale is the dominant resolved component in this controlled adaptation. Restricting both blocks to a single 1×1 scale reduces fused reversed-color accuracy from 57.27% to 13.46%, corresponding to a paired decrease of 43.81 percentage points. In the full configuration, the learned fusion places mean weights of 0.9425 on block 2 and 0.0575 on block 1. Final-local accuracy is 58.76%, only 1.48 points above fusion, indicating that most of the robustness present in the final local decision is retained after logit fusion.

Removing cross-channel probes decreases fused accuracy by 5.32 points, whereas removing feature alignment changes it by only 0.71 points in the opposite direction. Neither paired contrast is statistically resolved under the five-seed analysis. Among the ablations tested here, multiscale spatial pooling is therefore the only component with a clearly resolved contribution to reversed-color performance.

E PAIRED-COLOR REPRESENTATION AND PREDICTION DIAGNOSTICS

The following analyses connect spatial goodness to the structure of learned features and to changes in individual predictions. They use primary-protocol reruns with red and green variants of 2,000 base test images per seed, following the paired-color evaluation in Appendix B.4. Binary targets remain digits 0–4 versus 5–9. Representation retrieval, prediction transitions, and selected cases examine different aspects of the same comparison.

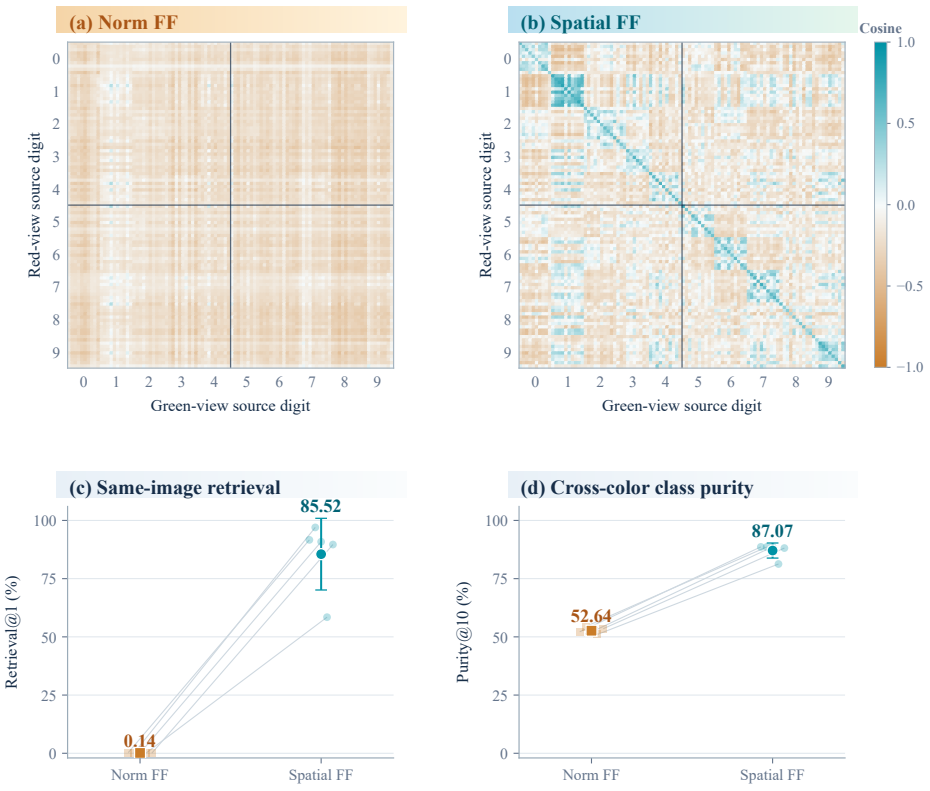

Figure A3: **Cross-color representation structure and retrieval.** (a,b) Cosine-similarity maps for red and green views of the same 100 seed-0 test images, with ten images per original digit in identical order for both methods. Dividing lines separate the two binary classes; both maps share one cosine scale. (c) Same-image retrieval@1 from a 2,000-image green-view gallery for each red-view query. (d) Binary-class purity among the ten nearest green-view neighbors, excluding the paired image. Quantitative metrics use all 2,000 queries per seed. Markers and error bars show five-seed means and sample standard deviations; faint points and lines show paired seed results.

E.1 CROSS-COLOR REPRESENTATION STRUCTURE

The representation analysis compares red-view queries with a green-view gallery. The reported feature space is standardized and candidate-safe, with 3,136 dimensions. Same-image retrieval asks whether the nearest cross-color neighbor is the query's paired base image. Binary-class purity measures the fraction of the ten nearest neighbors with the same target after excluding that paired image.

Spatial FF reaches 85.52% same-image retrieval, compared with 0.14% for norm FF, and its class purity is 87.07% versus 52.64% (Figure A3). The retrieval difference identifies stronger cross-color instance correspondence in this feature space. The purity difference extends the comparison beyond paired self-matches. The heatmaps display a fixed descriptive subset, while both quantitative

metrics use all queries. These representation measures complement the direct readout intervention in Figure A1.

E.2 Prediction transitions across original digits

Joint correctness identifies where the accuracy gain comes from. Across all paired test variants, 41.37% change from a wrong norm-FF prediction to a correct spatial-FF prediction, while 0.56% change in the opposite direction. The remaining 49.67% stay correct and 8.40% stay wrong. The correction-minus-degradation difference gives the 40.81-point net improvement in Figure A4.

The correction fraction ranges from 35.32% to 45.74% across original digits, while degradation ranges from 0.09% to 1.05%. Improvements therefore occur across all ten digit groups rather than being concentrated in a small subset. Each rate uses every evaluated variant in its seed or seed/digit group as the denominator. In particular, 41.37% is a share of all variants, not a conditional recovery rate among baseline errors.

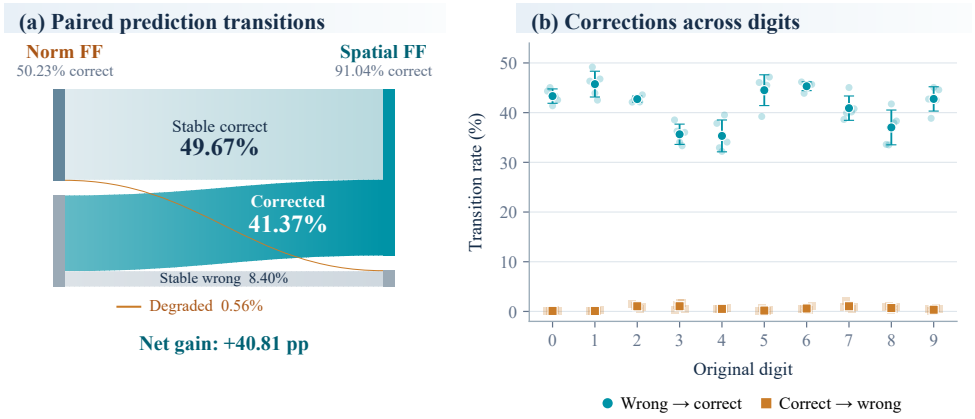

Figure A4: **Paired prediction transitions across original digits.** Each of 2,000 base images per seed is evaluated in red and green. (a) Flows from norm FF to spatial FF have widths proportional to five-seed mean joint transition rates: stable correct, 49.67%; corrected, 41.37%; degraded, 0.56%; and stable wrong, 8.40%. Flow widths show means without uncertainty intervals. (b) Correction and degradation rates by original digit, with five-seed means and sample standard deviations. Both panels use all variants in the relevant seed or seed/digit group as denominators.

E.3 Selected color-swap cases

Figure A5 shows eight seed-0 cases in which norm FF predicts class 0 for the red view and class 1 for the green view. Spatial FF predicts the true binary class for both colors. The native score margins make the transition explicit: one norm-FF margin is negative in each pair, while both spatial-FF margins are positive. Their magnitudes are interpreted within each model because the score scales differ. The cases illustrate individual corrections; the frequencies and digit coverage are quantified in Figure A4.

F Concept Audit and Targeted Cue Diversification

F.1 Concept representation and sparse rules

The concept vocabulary gives the audit a fixed view of color and coarse location. It contains the mean of each input channel over five regions. These regions are the full image and its four 14×14 quadrants. The resulting 15 concepts have direct channel and region names.

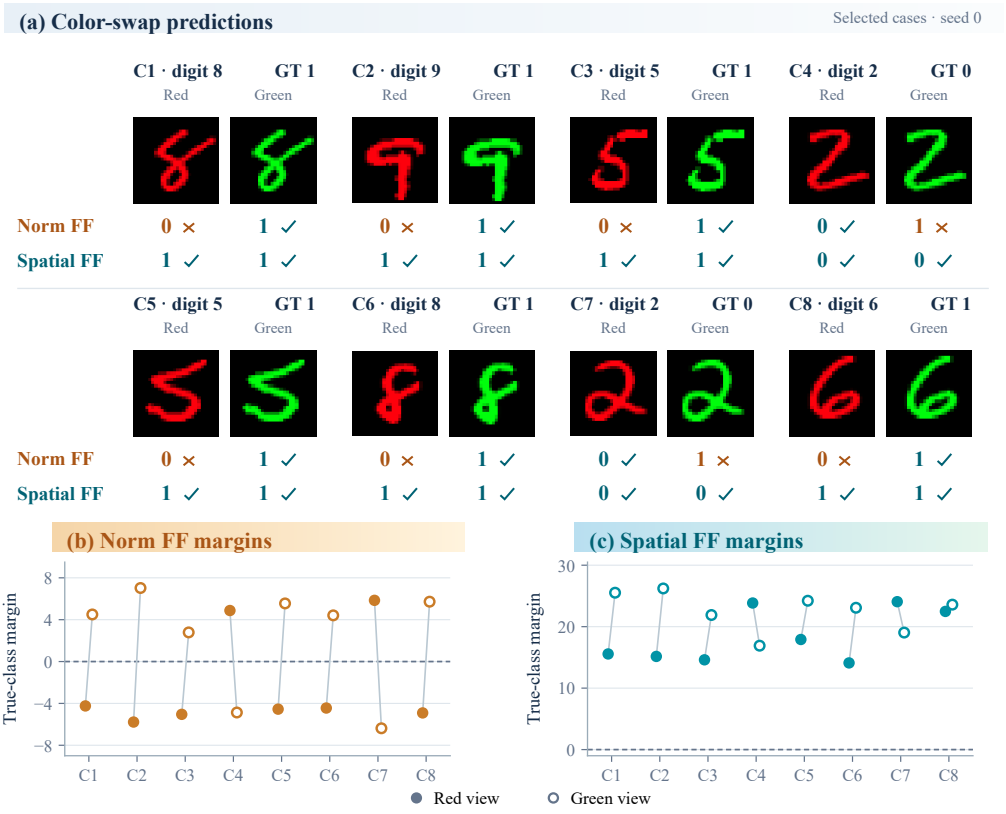

Figure A5: **Selected prediction corrections under paired color swaps.** (a) Eight seed-0 examples selected by a deterministic test-margin criterion; the panel illustrates individual corrections rather than their frequency. C1–C8 correspond to base-image IDs 5188, 4480, 3863, 2439, 5206, 6776, 2503, and 384. Each pair shows the same grayscale digit under red and green assignments while retaining the 0.05-intensity blue reference. GT 0 denotes digits 0–4 and GT 1 denotes digits 5–9. (b,c) Native true-class margins (true-class score minus the alternative-class score). Filled and open markers denote red and green views, respectively; positive margins indicate correct predictions. Separate axes preserve the model-specific score scales.

The rule model standardizes each concept by its training mean and standard deviation. Class-specific k-means centers initialize four rules per class. Each rule learns a center, a positive feature scale, and softmax-normalized feature weights. Its log firing strength is a weighted sum of Gaussian log memberships. A log-sum-exp over rule strengths produces the candidate class score.

The rule loss adds $\mathrm{softplus}(-s_y)$ for the positive candidate and $\mathrm{softplus}(s_{1-y})$ for the negative candidate. Their means are combined with a feature-weight entropy penalty of 0.015. Adam trains the rule model at learning rate 0.02 for 100 or 250 full-batch epochs in the reported configurations. The top three concepts name the evidence emphasized by the rules. Statistical triggering uses the full fixed vocabulary, keeping the tested statistic separate from learned concept ranking.

F.2 Reference-adjusted calibration

Calibration measures excess chromatic class separation over the shape-carrying reference. For each of 15 channel-region concepts, standardized separation is

$$d_j = \frac{|\mu_{1j} - \mu_{0j}|}{\sqrt{(\sigma_{1j}^2 + \sigma_{0j}^2)/2 + \epsilon}}. \tag{7}$$

Here μ_{yj} and σ^2_{yj} are the class-conditional mean and variance of concept j. The variance denominator uses $\epsilon = 10^{-8}$ for numerical stability. Red and green carry the assigned color cue; blue carries digit shape without that assignment. With color concepts $\mathcal{C}$ and their blue references $\mathcal{B}$, the score is

$$A = \max_{j \in \mathcal{C}} d_j - \max_{j \in \mathcal{B}} d_j. \tag{8}$$

Both maxima range over their fixed channel-region concepts. Subtracting the reference removes measured separation already carried by shape.

Each dataset is divided into disjoint stratified halves. The statistic uses the held-out half, with class proportions preserved by the split. Calibration uses seeds 100 through 119 in two null environments: replicated grayscale and label-independent color. This yields 40 environment scores. At level $\alpha = 0.05$, the chosen upper order statistic has rank $\lceil (40 + 1)(1 - \alpha) \rceil = 39$.

The recorded threshold is zero, and the trigger requires a strictly larger score. Calibration precedes evaluation on seeds 0 through 4. Evaluation outcomes from the FF classifiers do not set the threshold. The reported calibration run uses no permutation tests; its decision comes from the external null scores.

Table A7: Independent evaluation of the reference-adjusted audit. Each row contains five dataset seeds. The trigger threshold is zero.

Environment	Color agreement	Audit score	Triggered seeds
Replicated grayscale	n/a	0.000 ± 0.000	$0/5$
Colored	0.50	-0.250 ± 0.015	$0/5$
Colored	0.60	0.063 ± 0.023	$5/5$
Colored	0.70	0.449 ± 0.024	$5/5$
Colored	0.80	0.898 ± 0.013	$5/5$
Colored	0.90	1.727 ± 0.037	$5/5$
Colored	0.95	2.547 ± 0.058	$5/5$
Colored	1.00	4.492 ± 0.049	$5/5$

The score increases with color agreement across the tested environments. It stays at or below threshold in both null conditions. At every tested $\rho \geq 0.6$, it exceeds threshold in all five seeds. The reference channel supplies a known shape comparison because its semantics are fixed by data generation.

F.3 Targeted diversification

Targeted diversification changes the diagnosed cue while preserving images' binary labels. The repair experiment starts from deterministic training color, with 30,000 training and 5,000 test images. It retrains norm FF for 15 epochs after randomizing color. Always-randomized training provides a separate comparison using another seeded color assignment.

Table A8: Color diversification with norm FF. Values are five-seed percentages with sample standard deviations.

Training regime	Matched color	Reversed color
Deterministic original color	100.00	0.00
Audit-guided color randomization	84.37 ± 0.90	84.08 ± 1.21
Always randomize color	84.94 ± 1.44	83.50 ± 1.54

Randomization makes digit content useful across both test environments. The two interventions have similar mean accuracy in the five-seed repair experiment. These means provide a descriptive comparison, without an equivalence test. The independent calibration experiment measures whether the reference-adjusted statistic selects colored and null environments. The repair experiment measures the effect of diversifying the selected color cue.

G Natural-Image Readouts and Local Adaptation

G.1 Waterbirds splits and group metrics

Waterbirds pairs bird classes with correlated land or water backgrounds. The official split contains 4,795 training, 1,199 validation, and 5,794 test images. The reporting groups combine the binary bird label y with background b as $g = 2y + b$. Average accuracy weights all test images equally. Worst-group accuracy is the minimum of the four group accuracies.

All natural-image models start from torchvision's default ImageNet-pretrained ResNet-18 weights. The associated deterministic transform resizes, center-crops to 224×224, and applies ImageNet normalization. Cached inputs or features preserve this common preprocessing. The pretrained representation contains information learned through global backpropagation; the subsequent training studies local adaptation or readout learning. Waterbirds is accessed through the WILDS software interface and uses its own official dataset splits.

G.2 Frozen representation diagnostic

The frozen diagnostic holds visual features fixed while changing the head. ResNet-18 features are extracted after layer 4 at $512 \times 7 \times 7$ resolution. They are cached in half precision and converted to single precision for head training. The backbone remains in evaluation mode throughout feature extraction.

The FF heads append the candidate plane after feature extraction. A 1×1 convolution maps the resulting 513 channels to 64 channels, followed by ReLU. FF-Pooled averages squared activations. FF-Spatial instead applies a scalar linear head over the $64 \times 7 \times 7$ map. BP-GAP reads 512 global averages, while BP-Spatial reads all $512 \times 7 \times 7$ features.

Every head uses AdamW for 30 epochs, with learning rate 0.001, weight decay 0.0001, and batch size 128. The best validation worst-group checkpoint is restored before test evaluation. GroupDRO-Spatial uses the BP-Spatial architecture and four adaptive group weights. Its group-weight step size is 0.01. Training group labels enter GroupDRO's loss; validation group labels enter selection for every head.

Table A9: Waterbirds with one frozen ImageNet ResNet-18 representation. All heads use validation worst-group selection. GroupDRO also uses training groups. Values are five-seed percentages.

Head	Average accuracy	Worst-group accuracy
FF-Pooled	81.20 ± 0.29	51.71 ± 2.17
FF-Spatial	82.34 ± 1.00	58.50 ± 2.56
BP-GAP	80.21 ± 0.32	57.73 ± 1.91
BP-Spatial	80.74 ± 0.96	60.29 ± 4.01
GroupDRO-Spatial	84.91 ± 3.14	78.53 ± 2.61

Spatial FF raises worst-group accuracy by 6.79 paired points on the fixed representation. The difference between the two BP heads remains unresolved. The spatial FF and spatial BP comparison is also inconclusive. The clearest supported effect in this diagnostic is the readout change within FF.

G.3 Updating the representation from pixels

The pixel-based comparison trains four local stages of ResNet-18. Stage one contains the stem and the first residual stage. The remaining stages contain residual layers 2, 3, and 4. Their output widths are 64, 128, 256, and 512. Each stage trains for five epochs before its selected state is frozen.

CwComp-Pooled divides each stage's channels evenly between the two classes. It averages squared activations within each class group to form the logits. CwComp-Spatial retains a 4×4 grid of squared activations within each group. A class-specific linear head maps that grid to its local logit. The convolutional stage and its local head update together using cross-entropy.

Local optimization uses Adam with learning rate 0.0001, weight decay 0.0001, and batch size 64. Earlier stages remain frozen and in evaluation mode during each new stage's updates. Gradients stop at the active stage's input. Validation worst-group accuracy selects each stage in the main comparison. The final local head provides the local learner's prediction.

The global references update the full pretrained network from the same transformed pixels. They use AdamW for 20 epochs, learning rate 0.00001, weight decay 0.0001, and batch size 64. Global GroupDRO also updates four group weights with step size 0.01. Validation worst-group accuracy selects checkpoints for both global references. Selected states are restored before a single test evaluation per method and seed.

Validation-only development runs examined local learning rates 0.00003, 0.0003, and 0.001 on seed 0. They also examined unweighted sums of stage logits. These choices did not improve the selected spatial configuration. Test data were not evaluated in these development runs.

H Fusion Objectives and Access to Group Information

H.1 Detached local-logit fusion

Fusion asks whether predictions from several local stages contain complementary class evidence. Each frozen stage produces two logits, giving eight features per image. Training-set means and standard deviations standardize these features before concatenation. A linear 8-to-2 classifier maps them to the final prediction. All stages and local heads remain detached throughout fusion training.

The fusion classifier uses AdamW for 100 epochs, learning rate 0.01, weight decay 0.0001, and batch size 256. ERM minimizes cross-entropy. GroupDRO replaces it with a weighted sum of the four group losses, using weight-update step size 0.01. The main ERM and GroupDRO fusion experiments use validation worst-group checkpoint selection.

Table A10: Group information used in Waterbirds fusion protocols. "Average" and "Worst group" denote the validation selection metric. FG denotes fully groupless training and selection.

Protocol	Local selection	Fusion loss uses groups	Fusion selection
LocalFusion-ERM	Worst group	No	Worst group
LocalFusion-GroupDRO	Worst group	Yes	Worst group
ERM-AvgSelect	Worst group	No	Average
JTT	Worst group	No	Average
CVaR-20	Worst group	No	Average
FG ERM fusion	Average	No	Average
FG JTT fusion	Average	No	Average

H.2 Error-set retraining and CVaR

JTT first identifies training examples that are misclassified by an initial ERM fusion head. The initial head is trained for 20 epochs, after which its training errors define a fixed error set. A fresh fusion head is then trained for 100 epochs, assigning weight 20 to examples in that set. Both stages use only class labels, and the final checkpoint is selected by validation average accuracy. Thus, the reweighting signal is derived from model errors rather than from training-group annotations.

CVaR-20 provides a comparison based on generic high-loss emphasis. Within each minibatch, it averages the largest 20% of per-example cross-entropy losses. All remaining optimizer and checkpoint-selection settings match the average-selected ERM and JTT protocols. This comparison tests whether emphasizing currently high-loss examples produces the same behavior as retraining on a fixed error set.

We evaluate two levels of group-information access. In the deployment-only setting, all fusion objectives and fusion checkpoint selection avoid group labels, but the underlying local checkpoints were previously selected using validation worst-group accuracy. In the fully groupless setting, local

stages are instead selected by validation average accuracy, and all fusion checkpoints are also selected by validation average accuracy. Group labels are then used only after model selection for test-time reporting.

Table A11: Waterbirds group-information ablations. FG removes group labels from local-stage and fusion selection. Values are five-seed percentages with sample standard deviations.

Method	Average accuracy	Worst-group accuracy
ERM-AvgSelect	72.54 ± 0.74	24.14 ± 2.60
JTT	77.27 ± 1.92	31.37 ± 2.12
CVaR-20	73.02 ± 0.82	23.36 ± 3.33
FG Spatial-local	73.96 ± 3.02	22.18 ± 3.17
FG ERM fusion	71.95 ± 1.56	23.49 ± 2.83
FG JTT fusion	77.41 ± 2.31	28.50 ± 5.46

JTT reaches 31.37% worst-group accuracy with group-selected local stages. It improves on the corresponding final spatial head by 6.82 paired points. Average-selected ERM and CVaR remain near that local head's accuracy. The error-set intervention therefore accounts for the supported deployment-only gain.

Fully groupless JTT retains 28.50% worst-group accuracy. Its gain over the corresponding final local head is positive in every seed, with a mean of 6.32 points. The paired test remains inconclusive. The lower mean and larger variation show how upstream selection affects the information available to fusion.

H.3 CelebA transfer

CelebA tests the fully groupless protocol on a second natural-image task (Liu et al., 2015). It uses the official identity-disjoint splits with 162,770 training, 19,867 validation, and 19,962 test images. The target is the *Blond_Hair* attribute. The *Male* attribute defines reporting groups after model selection. It is excluded from training, sampling, weighting, hyperparameter choice, and checkpoint selection.

The pretrained ResNet-18, four local stages, and spatial heads follow the fully groupless Waterbirds configuration. Each local stage trains for five epochs with batch size 1,024. Fusion uses batch size 2,048 and the same 20-epoch identification, weight-20, and 100-epoch retraining schedule. All local and fusion checkpoints use validation average accuracy. The four test groups contain 9,767, 2,480, 7,535, and 180 images.

Table A12: Fully groupless CelebA transfer. Model selection uses validation average accuracy. Values are five-seed percentages with sample standard deviations.

Method	Average accuracy	Worst-group accuracy
Spatial-local	95.82 ± 0.07	44.33 ± 1.77
ERM fusion	95.88 ± 0.06	41.78 ± 1.54
JTT fusion	95.48 ± 0.28	42.78 ± 9.02

High average accuracy coexists with much lower accuracy in the least accurate group. JTT changes worst-group accuracy by -1.56 paired points relative to the final spatial head. Its sample standard deviation rises to 9.02 points. The test does not support a JTT improvement on this task, despite the Waterbirds gain.

Table A13: Eight-group worst accuracy in the Waterbirds multi-shortcut stress test over five seeds. "Aligned," "random," and "reversed" denote test patch–label correlations 0.9, 0.5, and 0.1. Values are percentages.

Method	Aligned	Random	Reversed
CwComp-Pooled	0.00 ± 0.00	0.38 ± 0.27	0.31 ± 0.29
CwComp-Spatial	5.35 ± 4.25	6.16 ± 3.61	5.94 ± 3.29
BP-ERM	25.09 ± 4.80	28.27 ± 5.54	28.00 ± 6.45

I Joint Cues, Input Location, and Audit Specificity

I.1 Waterbirds with two simultaneous shortcuts

The joint-cue experiment adds a controllable patch to Waterbirds' natural background correlation. A 20×20 patch occupies a fixed corner after ImageNet normalization. Its values correspond to normalized pure red for cue 0 and pure green for cue 1. Independent seeded Bernoulli draws make the patch agree with the class in 90% of training and validation examples.

The model trains once per seed using the main pixel-based protocol. Three held-out environments change patch agreement to 0.9, 0.5, or 0.1. The natural backgrounds remain unchanged. Validation selection uses the original four label-background groups. Evaluation also reports eight joint groups combining label, background, and patch assignment.

Every joint test group is nonempty. Under aligned patches, the smallest groups contain roughly 60 examples per seed. The primary metric takes the minimum accuracy across all eight groups. This metric exposes failures that can be hidden when patch assignments are averaged within a background group.

Spatial local goodness improves this minimum over pooled goodness in all three environments and all seeds. The gains range from 5.35 to 5.78 percentage points. Absolute joint worst-group accuracy remains close to zero for both local learners, as shown in Table A13. BP-ERM exceeds the spatial local model by 19.74 to 22.11 points. Preserving spatial evidence improves the local learner while leaving strong dependence on joint cue combinations.

I.2 A location cue with equal patch energy

The location experiment moves the cue from channel identity to image position. A white 6×6 patch occupies the top-left or bottom-right corner of balanced CIFAR-10 images. Patch size and intensity remain fixed. Its location agrees with the binary class with probability 0.9 during training. Reversed-cue testing changes this agreement to 0.1.

Table A14: CIFAR-10 location-cue results. Values are five-seed percentages with sample standard deviations. The cue changes patch position while keeping its area and intensity fixed.

Learner	Matched cue	Random cue	Reversed cue
FF-Pooled	84.91 ± 1.32	59.07 ± 2.69	33.54 ± 5.02
FF-Spatial	91.32 ± 0.53	65.27 ± 1.63	38.80 ± 2.57

Spatial FF has a mean reversed-cue gain of 5.26 points, with one seed favoring pooled FF. The paired comparison remains inconclusive. This result separates input-location behavior from the exact permutation invariance of the goodness functional. Moving an input patch can alter convolutional activations as well as their locations.

I.3 What the audit must distinguish

Cross-class concept overlap alone is insufficient to distinguish a shared shortcut from shared task structure in this protocol. A Jaccard trigger compares the highest-ranked concepts from the two

class-specific rule banks. In the evaluated development seed, colored data yield Jaccard overlap 0.2 and do not trigger, whereas replicated grayscale yields overlap 1.0 and does trigger. Distinct red and green concept names can obscure a shared color mechanism, while replicated channels can create high overlap in the absence of a color shortcut.

Raw class separation is also sensitive to legitimate digit structure. Thresholding the largest unadjusted concept effect can therefore trigger on clean examples when the shape-carrying blue reference separates the two digit groups. The reference-adjusted statistic instead subtracts this measured blue-channel separation before thresholding. Its independent null calibration and evaluation are reported in Appendix F.

The classifier comparison requires a similar distinction. Comparing FF and BP only under global pooling leaves both learners strongly dependent on the shortcut. Adding spatial versions changes the readout geometry under both training rules and reveals a common spatial-reduction effect. These controls make readout geometry the shared manipulated factor, while leaving the FF-versus-BP training rule as a separate comparison.

J Statistical Tests and Prespecified Decision Rules

J.1 Paired-seed inference

Reported means and standard deviations summarize five independently seeded dataset–model runs. Method comparisons are paired by seed and data construction. For paired seed-level differences $d_1, \ldots, d_5$, let $\bar{d}$ denote their mean and s_d their sample standard deviation. We report two-sided 95% Student-t intervals with four degrees of freedom:

$$\bar{d} \pm t_{0.975,4} \frac{s_d}{\sqrt{5}}. \tag{9}$$

The tables below report these 95% intervals in percentage points. Tests are paired and two-sided. The intervals describe variation across seeds rather than independent binomial sampling of each test image. Reported p values are unadjusted across contrasts.

Table A15: Paired geometry contrasts. Rows report accuracy in the corresponding shifted test environment: reversed color for color-cue experiments and reversed cue for the location-cue experiment. All intervals are two-sided 95% Student-t intervals over five paired seeds.

Protocol	Contrast	Mean	Interval
Primary MNIST	Spatial FF minus norm FF	74.19	$[72.57, 75.81]$
Primary MNIST	Spatial FF minus spatial BP	5.74	$[-0.21, 11.69]$
Capacity	Spatial minus invariant MLP	56.98	$[53.55, 60.41]$
Capacity	Stable minus randomized spatial	66.40	$[61.44, 71.36]$
Capacity	Spatial minus pooled	75.63	$[72.37, 78.90]$
Colored CIFAR-10	Spatial minus pooled	10.32	$[7.91, 12.73]$
Location CIFAR-10	Spatial minus pooled	5.26	$[-3.47, 13.99]$

The primary Spatial-FF versus Norm-FF contrast and both capacity controls have strictly positive intervals. The Spatial-FF versus Spatial-BP interval crosses zero, so the primary experiment does not resolve a difference between the two spatial training rules. Colored CIFAR-10 retains a positive spatial-versus-pooled contrast ($p = 0.000285$), whereas the location-cue comparison remains unresolved ($p = 0.170$). Together, these results distinguish the strongly supported color-shift effect from the weaker evidence in the input-location experiment.

Every spatial-versus-pooled interval in the matched sweep is positive. The comparison between spatial training rules varies with correlation. Its $\rho = 0.95$ interval crosses zero, with $p = 0.059$. The sweep therefore consistently supports spatial preservation without establishing a general ordering between local and global training.

ASGE's two finer spatial configurations have positive local-minus-fusion differences in every seed. Its two configurations with a 1×1 final grid have intervals crossing zero. BiCovG's scale contrast

Table A16: Matched correlation-sweep contrasts in reversed-color accuracy. Entries give mean paired gain and its 95% interval in percentage points.

ρ	Spatial FF minus norm FF	Spatial BP minus BP-GAP	Spatial FF minus spatial BP
0.60	25.83 [19.55, 32.12]	10.90 [7.33, 14.47]	1.72 [0.07, 3.37]
0.70	43.92 [42.60, 45.23]	15.57 [12.11, 19.04]	1.56 [0.38, 2.74]
0.80	63.06 [61.41, 64.72]	35.19 [25.52, 44.86]	2.50 [0.55, 4.44]
0.95	76.04 [73.92, 78.17]	69.95 [65.28, 74.62]	5.89 [−0.38, 12.16]

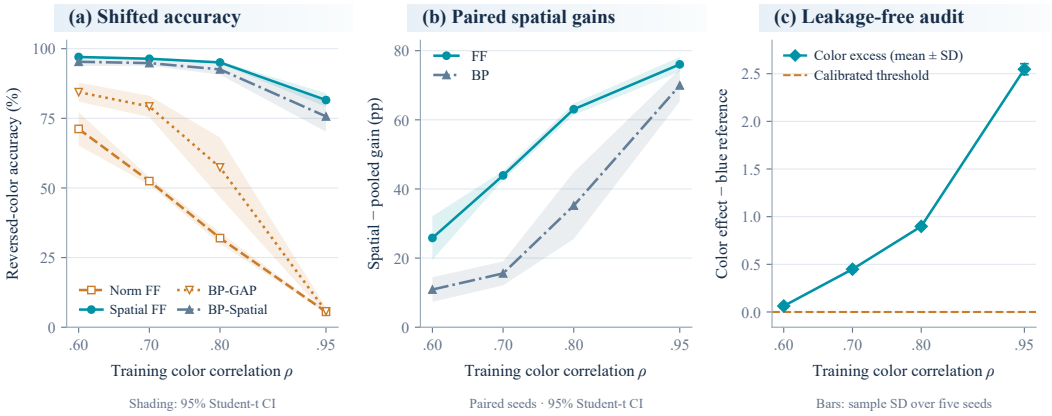

Figure A6: Matched correlation sweep with 30,000 training images, 5,000 test images, 15 epochs, and five seeds. Shading gives two-sided 95% Student-t intervals for learner means in (a) and paired spatial-minus-pooled gains in (b). Bars in panel (c) give sample standard deviations of audit scores. The separate seed-0 deterministic-color stress test is excluded.

Table A17: Paired contrasts for the adapted local objectives. Values compare reversed-color accuracy in percentage points.

Contrast	Mean	95% interval	p
CwComp minus pooled FF	59.64	[54.07, 65.21]	0.000008
Spatial FF minus CwComp	16.25	[12.58, 19.91]	0.000250
ASGE 7/4: local minus fusion	67.36	[63.99, 70.72]	< 0.000001
ASGE 4/2: local minus fusion	60.44	[57.61, 63.26]	< 0.000001
ASGE 2/1: local minus fusion	0.38	[−0.38, 1.14]	0.239
ASGE 1/1: local minus fusion	3.27	[−0.59, 7.14]	0.078
BiCovG: local minus fusion	1.48	[0.46, 2.51]	0.016
BiCovG: full minus single scale	43.81	[37.25, 50.38]	0.000050
BiCovG: full minus no probes	5.32	[−6.06, 16.71]	0.264
BiCovG: full minus no alignment	−0.71	[−4.16, 2.73]	0.597

is positive in all seeds, while its probe and alignment contrasts remain unresolved. The inference matches the component-level patterns in Appendix D.

Table A18: Paired natural-image contrasts in worst-group accuracy. FG denotes fully groupless selection. Values are percentage points.

Contrast	Mean	95% interval	p
Frozen: spatial FF minus pooled FF	6.79	[2.35, 11.23]	0.013
Frozen: spatial BP minus BP-GAP	2.57	[−3.92, 9.06]	0.334
Frozen: spatial FF minus spatial BP	−1.79	[−7.27, 3.69]	0.416
Waterbirds: spatial minus pooled local	14.49	[9.70, 19.27]	0.001
Waterbirds: ERM fusion minus spatial local	−0.12	[−7.66, 7.41]	0.966
Waterbirds: JTT minus spatial local	6.82	[0.51, 13.13]	0.040
Waterbirds: JTT minus ERM fusion	6.95	[1.59, 12.30]	0.023
Waterbirds FG: JTT minus spatial local	6.32	[−0.86, 13.51]	0.071
Waterbirds FG: JTT minus ERM fusion	5.02	[−0.69, 10.73]	0.071
GroupDRO fusion minus spatial local	26.88	[19.18, 34.58]	0.001
GroupDRO fusion minus BP-ERM	−2.34	[−10.59, 5.91]	0.474
GroupDRO fusion minus BP-GroupDRO	−28.04	[−33.86, −22.22]	0.000180
CelebA FG: JTT minus spatial local	−1.56	[−13.01, 9.90]	0.725
CelebA FG: JTT minus ERM fusion	1.00	[−9.07, 11.07]	0.796

Ordinary Waterbirds fusion has an interval centered near zero relative to the final local head. Deployment-only JTT has a positive interval, while fully groupless JTT remains inconclusive despite positive gains in every seed. GroupDRO fusion and global ERM also have an interval crossing zero. CelebA provides no supported JTT gain over either comparison head.

Table A19: Waterbirds joint-cue contrasts in eight-group worst accuracy. Values are percentage points with paired 95% intervals.

Patch environment	Spatial minus pooled local	Spatial local minus BP-ERM
Aligned	5.35 [0.08, 10.62]	−19.74 [−27.60, −11.87]
Random	5.78 [1.44, 10.13]	−22.11 [−28.02, −16.20]
Reversed	5.63 [1.67, 9.58]	−22.06 [−28.01, −16.10]

J.2 Prespecified fusion criteria

The fusion studies distinguish a detectable change from a material robustness target. Waterbirds deployment-only fusion required mean worst-group accuracy of at least 40% and a gain of at least ten points. The comparison reference was CwComp-Spatial at 24.55%. JTT reached 31.37% and gained 6.82 points, satisfying neither material target. Its positive paired contrast supports a smaller improvement.

The fully groupless study retained the same material target and added a group-removal preservation rule. Preservation required at least 28% worst-group accuracy and a loss of at most five points from deployment-only JTT. Its 28.50% result is 2.87 points below 31.37%, satisfying both preservation conditions. The separate criterion for a positive local-improvement interval was not met.

CelebA required JTT worst-group accuracy of at least 60%, a five-point local-head gain, and a positive paired interval. The observed 42.78% accuracy, −1.56-point difference, and zero-crossing interval meet none of these conditions. Gains vary in sign across seeds, leaving the five-positive-seed directional criterion unmet. These thresholds were fixed before the corresponding test outcomes were inspected.

K Code and Reproduction Package

The supplementary package provides an anonymous standalone implementation of the rule-based goodness feasibility study. It uses public scikit-learn datasets and requires no separate data download. The script compares a BP-trained MLP, squared-goodness FF, and dense and sparse fuzzy-rule FF variants under fixed random seeds and standardized preprocessing. It reports accuracy, macro-F1, negative log-likelihood, calibration error, training time, and interpretable fuzzy-rule statistics, and exports per-seed results, aggregate summaries, rules, and run metadata in CSV and JSON formats. The package is intended as a compact reference implementation of the rule-based goodness component; full protocols for the image-based experiments are provided in the corresponding appendices.