WEAK LEARNING FAILURE ALONE IS NOT GRADIENT STARVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient starvation describes competition in which learning one predictive feature suppresses learning another. Yet a weak response alone cannot establish suppression: the same feature may remain poorly learned without the competing cue. We develop a controlled approach using matched training runs with an informative cue, an uninformative cue, and no cue. Our criterion requires both reference runs to reach a learning target within a fixed budget and measures the informative-cue deficit at each reference's first target crossing and across training. Across four model families, comparisons with uninformative cues show why negative differences alone are insufficient evidence. We derive an exact gradient-flow decomposition that tracks how differences in output loss gradients and response sensitivities, together with weight decay, change weak-feature learning. In a noiseless two-layer linear model under unregularized logistic gradient flow, we prove that a perfectly reliable cue can sustain a relative learning deficit indefinitely, even as both runs continue learning. We also rigorously verify finite-time suppression bounds against cue-free training in small networks through validated numerical integration. In CIFAR-10, removing or randomizing the cue during training improves cue-free accuracy by about 7.6 percentage points relative to retaining it, measured using a fresh linear classifier fitted to frozen features. On pretrained Waterbirds models, poor worst-group accuracy coexists with small measured core-feature deficits. Together, these results distinguish cue-induced suppression from weak learning shared across training conditions and show why poor robustness need not imply a substantial feature-learning deficit. Code is available at `https://anonymous.4open.science/r/gradient_starvation-C646/`.

1 INTRODUCTION

A neural network can predict a label from an easy cue while learning another useful feature poorly. A bird classifier, for example, may rely heavily on the background. But poor performance when that background changes does not tell us how much the model learned about the bird. Even a small response to the bird itself leaves an unanswered question: would the same training procedure have learned more without the background cue?

This distinction matters when diagnosing *gradient starvation*, in which learning one predictive feature suppresses learning another (Pezeshki et al., 2021). A weak feature may be difficult to learn at the available training budget, whether or not a competing cue is present. Conversely, a model may learn useful features but give them too little weight when making predictions (Kirichenko et al., 2023). These explanations call for different comparisons. We use *weak feature* for the predictive feature whose learning we want to assess, and *cue* for the competing input signal.

We compare three training runs that share their initialization, data order, and other training choices. One receives an informative cue, one receives a cue that is independent of the label, and one receives no cue. The run without a cue asks whether the training procedure can learn the weak feature within the available time. The uninformative-cue run asks whether a difference depends on the cue's label information, rather than merely its presence. We measure all three on the same inputs with the cue removed. Figure 1 shows why observing a low response is insufficient: the same response can lead to different conclusions depending on what the two matched reference runs learn within the same training budget.

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.

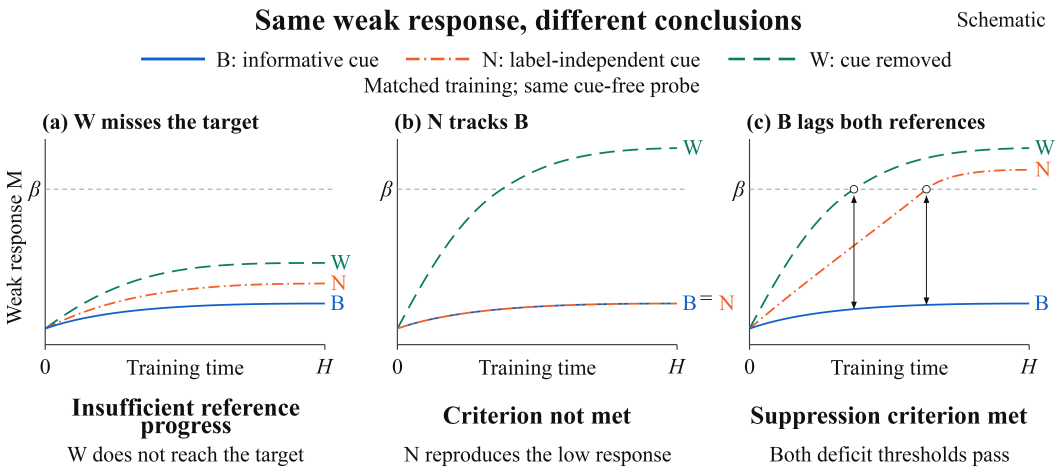


Figure 1: Schematic matched comparisons: **(a)** W misses the target; **(b)** label-independent N tracks B; **(c)** both references reach the target and B meets the relative first-hit (arrows) and normalized signed integrated-deficit thresholds against each (Definition 1). Curves are illustrative, not measurements.

The comparison must also distinguish a substantial learning deficit from a small or brief fluctuation. Our criterion therefore requires each comparison run to reach a chosen response level. The informative-cue run must fall behind by a minimum amount when that level is first reached and when the difference is averaged over training. If the model trained without the cue never reaches the target, the comparison is inconclusive at that budget. The criterion makes these choices explicit; it is not a universal definition of starvation.

The paper makes three contributions. First, it gives a practical way to compare weak-feature learning while checking both reference progress and the effect of an uninformative cue. Second, it explains several features of these comparisons mathematically: an uninformative cue can change early learning, and a perfectly reliable cue can cause a lasting relative deficit in a linear model. Third, it connects these comparisons to interventions and robustness. Removing or randomizing a CIFAR-10 cue partly reverses the measured deficit, whereas pretrained Waterbirds models show that poor worst-group accuracy need not imply a large deficit in the chosen core-feature response.

2 Related Work

Why models prefer easy cues. Shortcut learning describes a model's reliance on features that work well in training but can fail when the data distribution changes (Geirhos et al., 2020). Studies of simplicity bias show that networks can favor simpler predictive features (Shah et al., 2020), while experiments on feature representations and cue choice show that this preference depends on the data, architecture, and training procedure (Hermann & Lampinen, 2020; Scimeca et al., 2022). Hermann et al. (2024) distinguish how predictive a feature is from how easily a model can extract it. These studies explain why a cue may dominate prediction. Our question concerns the effect on another feature: how much less does the model learn about that feature when the cue is available?

Gradient starvation and learning dynamics. Pezeshki et al. (2021) study this interaction directly. They define starvation through the effect of increasing one feature's strength on another feature's response at the optimum of a linearized, ridge-regularized model, and propose spectral decoupling to reduce the interaction. Puli et al. (2023) relate shortcut learning to cross-entropy gradients and max-margin behavior. More generally, implicit-bias results describe the directions selected by logistic training in linear and deep linear models (Soudry et al., 2018; Ji & Telgarsky, 2019), and relate homogeneous-network training to constrained stationary points (Lyu & Li, 2020). In an infinite-width kernel regime with random initialization, linear RNNs have a bias toward short memory (Emami et al., 2021). We build on these accounts by following the difference between training runs over a specified budget. This lets us separate an early delay from a lasting relative deficit. Our loss comparisons concern the tested protocols and do not contradict guarantees that rely on different assumptions.

Learning a feature and using it. A model's predictions do not fully describe the information in its representation. Kirichenko et al. (2023) show that models relying on spurious features can still learn useful core features; retraining the final layer can then improve robustness. Methods such as group distributionally robust optimization, Just Train Twice, and Learning from Failure address poor performance on groups where the usual correlation breaks (Sagawa et al., 2020; Liu et al., 2021; Nam et al., 2020). Their goal is improved prediction under these shifts. Here we ask a complementary question about feature learning, using the same weak-feature measurement across training conditions. This distinction is especially relevant to the pretrained Waterbirds experiment.

Together, these works motivate checking what the same training procedure learns with and without an informative cue. We add an uninformative-cue comparison because a cue can change optimization even when it carries no label information. The next section makes these comparisons precise by fixing what we measure, how long we train, and how large a deficit must be to count as suppression.

3 Comparing feature learning

3.1 Three training runs and one measurement

The three runs differ only in the cue. In the *informative-cue run* (B), the cue agrees with the label with reliability r. In the *cue-free run* (W), the cue is removed. In the *uninformative-cue run* (N), it has the same amplitude as in B but is independent of the label. The weak-feature inputs are shared, as are the initialization θ_0, data order, augmentation, optimizer, and training randomness. Thus B versus W measures the effect of including the informative cue, while B versus N compares an informative and an uninformative cue under the same protocol.

We assess weak-feature learning with a fixed scalar response $M(\theta)$, measured on inputs with the cue removed. Larger values mean a stronger weak-feature response. In the synthetic models, we compare the model's outputs on inputs that differ only in the weak feature. On CIFAR-10 and Waterbirds, we use class-balanced Brier skill, a score based on predicted class probabilities: uniform predictions score zero and perfect predictions score one. The exact measurements are in Appendix A. These responses measure particular aspects of learning; they need not capture every useful feature in a representation.

Write $M^X(t) = M(\theta_X(t))$ for run X at training time t. For two runs, the response difference and its rate of change are

$$\Delta_{X|Y}(t) = M^X(t) - M^Y(t), \qquad d_{X|Y}(t) = \dot{\Delta}_{X|Y}(t), \qquad \Delta_{X|Y}(0) = 0. \tag{1}$$

A negative difference means that X has learned less according to this response. It does not yet tell us whether the difference is large or persistent enough to matter.

3.2 Requiring reference progress and a substantial deficit

Before looking at the comparison, we choose a training budget H and a response target β above both zero and the initial response. A reference run must reach β within H. This requirement prevents us from attributing poor learning entirely to the cue when the same training procedure also makes little progress in the reference run.

We then check the deficit in two ways. First, the informative-cue run must be behind when the reference first reaches the target. Second, the accumulated deficit over training must exceed a chosen threshold. The second condition discounts brief differences; because it uses a signed area, an earlier advantage can offset a later deficit.

Definition 1 (Suppression criterion)**.** Fix a continuous response M, a target $\beta > \max\{0, M(\theta_0)\}$, a budget $H > 0$, and thresholds $\epsilon \in (0, 1)$ and $\alpha > 0$. For reference run Y, let

$$\tau_Y = \inf\{t \in [0, H] : M^Y(t) \geq \beta\}, \qquad a_{X|Y}(H) = \frac{\int_0^H (M^Y(t) - M^X(t))\,\mathrm{d}t}{\int_0^H [M^Y(t)]_+\,\mathrm{d}t},$$

where $\inf \emptyset = \infty$ and $[z]_+ = \max\{z, 0\}$. We write $\mathrm{MS}_{\epsilon,\alpha}(X \mid Y)$ when all three conditions hold:

1. The reference reaches the target: $\tau_Y \leq H$.

2. The deficit at that time is at least $\epsilon M^Y(\tau_Y)$: $\Delta_{X|Y}(\tau_Y) \leq -\epsilon M^Y(\tau_Y)$.

3. The normalized accumulated deficit satisfies $a_{X|Y}(H) \geq \alpha$.

B meets the full suppression criterion when both $\mathrm{MS}_{\epsilon,\alpha}(\mathrm{B} \mid \mathrm{W})$ and $\mathrm{MS}_{\epsilon,\alpha}(\mathrm{B} \mid \mathrm{N})$ hold.

Figure 1 illustrates three outcomes of this rule. Continuity and the positive target ensure that the area denominator is positive whenever the reference reaches the target. If W does not reach it, we report insufficient reference progress. Otherwise, we report whether the full criterion is met. Failure of N to reach the target prevents the full criterion from being met. A result that does not meet the criterion does not establish that the cue had no effect.

3.3 Choosing the deficit threshold

We fix $\epsilon = 0.05$ and choose α separately for each model family using runs with an uninformative cue. Specifically, we apply the two-run condition $\mathrm{MS}_{\epsilon,\alpha}(\mathrm{N} \mid \mathrm{W})$ and choose the smallest value on a fixed grid for which at most 5% of these calibration comparisons meet the condition. We call this value $\alpha^\star$ and freeze it before evaluating separate confirmation seeds. We abbreviate $\mathrm{MS}_{\epsilon,\alpha}$ to MS_α when $\epsilon = 0.05$.

This is an empirical choice of threshold. It does not guarantee a 5% population error rate, and the N-versus-W check is only one component of the full three-run rule. For comparison, we also evaluate an *any-negative-gap rule*: the reference must reach the target, but any negative difference during training is enough. This rule has no minimum deficit requirement.

The response's zero, scale, and target are fixed before evaluation. Multiplying M and β by the same positive constant leaves the criterion unchanged; adding an offset generally does not. Appendix A gives the implementation on logged training points and the mapping to the original result-file identifiers.

4 How cues affect learning

The comparisons above raise three questions: why can an uninformative cue change learning, how does an informative cue change the weak-feature response, and can longer training remove the deficit? We summarize the answers here. Full statements and proofs are in Appendix B; their numerical checks are in Appendix C.

4.1 An uninformative cue can change early learning

Independence from the label does not make a cue irrelevant to optimization. In a noiseless linear RNN, we can calculate the initial rate of change of the weak-feature difference exactly (Theorem 1). The rate depends on both the model's initial responses and how its shared parameters couple the two features.

For an uninformative cue, reversing the sign of the initial output weights reverses this rate. If the initialization distribution is unchanged by that reversal and the rate is nonzero with probability one, it is negative for half of the initializations (Corollary 1). Thus a rule that accepts any early negative difference can respond even to a cue with no label information. This is a local statement about the initial rate, not a prediction of how often the finite-budget criterion will be met. It explains why we check the size and duration of the deficit as well as its sign.

4.2 Two sources of change in the weak response

A training example affects the weak feature through two quantities: the loss gradient at the model's output, and the way parameter changes alter the measured weak response. These quantities can both differ between runs. Under gradient flow, the continuous-time limit of gradient descent, their contributions to the rate of change of the response difference satisfy

$$\dot{\Delta}_{X|Y} = \mathrm{SW} + \mathrm{G} + \mathrm{WD}. \tag{2}$$

The term SW captures differences in output loss gradients while averaging the two runs' response sensitivities. The term G captures differences in those sensitivities while averaging the loss gradients. WD is the contribution from weight decay. Proposition 1 gives the exact formulas at differentiable states. The first term is related to the residual-weight explanation of starvation (Pezeshki et al., 2021); the second describes how the model's parameters affect the response.

This decomposition is an identity, not a separation into independent causal mechanisms. Both terms depend on the evolving parameters. Their integrated contributions can have opposite signs, so a contribution ratio can be negative or exceed one.

We also consider adding the cue directly to the output through a separate, zero-initialized weight, often called *late fusion*. At a differentiable initial state, the cue changes neither the predictions nor the weak-feature learning rate. A stronger result holds under unregularized logistic gradient flow when all label-signed inputs are identical, the cue is perfectly reliable, and the weak probe matches the training input. If the core satisfies $f(-x) = -f(x)$ and has a locally Lipschitz gradient, the cue can only slow progress along the same weak-learning path (Proposition 3). These assumptions exclude arbitrary noisy inputs. At later times the two runs can have different parameters, even though their initial rates agree.

4.3 What longer training can change

Consider a noiseless two-layer linear model trained by logistic gradient flow without weight decay. The cue has strength ρ, the weak feature has unit strength, and both occur at the same time in the equivalent zero-lag RNN. If the cue is perfectly reliable and neither run starts on the exceptional set where the training margin stays bounded, then

$$\frac{M_w^{\mathrm{B}}(t)}{M_w^{\mathrm{W}}(t)} \longrightarrow \frac{1}{1+\rho^2} \qquad \text{as } t \longrightarrow \infty. \tag{3}$$

Here M_w is the weak-feature response. Both runs continue learning, but the run with the cue keeps only a fraction of the cue-free response. The relative deficit therefore persists. The precise initialization conditions are in Theorem 2(a). The cue-free run eventually reaches any fixed positive target, although the required time grows exponentially with the target (Theorem 2(a)).

An imperfect cue gives a different prediction. If both agreeing and disagreeing examples are present and the logistic loss falls below $\log 2$ at some time, the weak coordinate becomes the limiting maximum-margin direction (Theorem 2(b)). That direction result does not establish equal learning rates between runs. We conjecture that the weak-response ratio then tends to one and the normalized accumulated deficit tends to zero (Conjecture 1). The experiments test finite-time behavior separately from this conjecture. Regularization or non-separable data can instead lead to finite limiting responses; Theorem 2(c) states the result for models linear in their parameters.

4.4 Rigorous finite-time bounds in small networks

A sampled training curve can miss behavior between evaluation times. For small smooth networks, we instead enclose the possible parameter trajectories over successive time intervals using validated numerical integration (Lohner, 1987; Nedialkov et al., 1999; Johansson, 2017). If every enclosure satisfies the required response inequality, the inequality holds throughout the interval (Theorem 3).

Across 1,000 sampled initializations for each of nine networks with 5 to 145 parameters, 41–51% satisfy $\Delta_{\mathrm{B|W}} \le -0.1M^{\mathrm{W}}$ throughout $[H/2, H]$, together with the requirement that W reaches its target. This is a relative suppression bound wherever $M^{\mathrm{W}} > 0$. The horizon is selected separately for each initialization from W's progress, so these frequencies describe that selection rule. These bounds compare B with W over the stated interval; they do not verify the N comparison, the deficit at the first target crossing, or the accumulated-deficit condition. Table 7 gives all results and Appendix B.6 explains the construction.

5 Experiments

Data, models, and measurements. We use four model families: linear RNNs (F-lin), tanh RNNs and GRUs (F-nl), MLPs learning an injected pattern on MNIST or Fashion-MNIST backgrounds

Table 1: Results on confirmation runs. The two N|W columns compare the uninformative-cue run N with the cue-free run W. The last column reports the full criterion among B runs whose corresponding W reaches the target. This last comparison uses the prespecified $\rho \leq 4, \ell \leq 4$ subset for F-lin and all conditions for the other families. Values are fractions, with counts in parentheses.

Family	Any negative difference, N|W	$\alpha^\star$	Minimum deficit condition, N|W	Full criterion B vs. W and N
F-lin	0.705	0.75	0.053 (34/640)	0.667 (192/288)
F-nl	0.997	0.25	0.069 (22/320)	0.997 (318/319)
F-e2r	0.306	0.05	0.053 (17/320)	0.980 (96/98)
F-cifar	1.000	0.15	0.000 (0/8)	1.000 (8/8)

(F-e2r), and a CNN on CIFAR-10 (F-cifar) (LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009). The injected-pattern experiments use labels independent of the background image, so they test learning of the added pattern rather than natural image semantics. Appendix D gives architectures, training settings, targets, and seed ranges.

In CIFAR-10, the cue is a ribbon appended to the image. Rp gives the correct class code with probability $p/100$ and a uniformly chosen wrong code otherwise. Thus R90 is the informative-cue run B, R10 is the uninformative-cue run N, and W0 is the cue-free run W. The suppression criterion uses the model's original prediction head and class-balanced Brier skill on images with the ribbon removed. We separately fit a fresh linear prediction head to frozen features to measure information accessible to a new classifier. We identify that measurement as *fresh-head accuracy*; it is not the response used by the criterion. Accuracy differences are in percentage points (pp), and reported intervals are 95%.

5.1 Checks with uninformative cues

A rule that counts any negative difference responds frequently to uninformative cues: 31–100% of confirmation comparisons meet that rule (Table 1). Requiring a minimum deficit sharply reduces these frequencies. At the thresholds selected on separate calibration runs, 5.3–6.9% of control comparisons meet the condition in the linear, nonlinear, and injected-pattern families. All three frequencies remain above the intended 5% calibration target.

For CIFAR, none of the eight controls meets the minimum-deficit condition, but the nominal upper confidence bound is 36.9%. Moreover, synthetic and injected-pattern records reuse seeds across conditions. The pooled intervals in Table 4 do not account for that dependence. These are descriptive control frequencies, not established error rates for the full criterion. The final column likewise reports how often the criterion is met, not its sensitivity against independently known starvation labels.

The full criterion also requires B to fall behind N. A post-hoc check counts how often this additional comparison changes the outcome. At the frozen thresholds, the N comparison removes only 3 of 617 positive B-versus-W results. Thus most of the reduction in these data comes from the deficit requirements and their calibration. The N comparison still asks a distinct question about label information. Exploratory threshold changes and the per-family counts are reported in Appendix E.

5.2 Training time and cue reliability

The reference's progress determines whether a weak response can be interpreted as suppression. Several models that miss the target at short budgets reach it with more training: the GRU by $H = 10$, the lag-8 linear RNN by $H = 25$, and patch CNNs on MNIST and Fashion-MNIST by $H = 100$ (Figure 2d). The earlier fixed-threshold B-versus-W condition then holds in 8/8, 32/32, and 62/64 records, respectively.

The smallest injected pattern gives the opposite example. At amplitude $a_1 = 0.001$, the cue-free model reaches the target in 0/64 records at tested budgets up to $H = 5000$. At the standard budget $H = 100$, the informative-cue model has a large cue response and a below-target weak response in 32/32 records per dataset. Yet its difference from the cue-free run accounts for only 0.003 and 0.017 of its shortfall on the two datasets. Poor weak-feature learning is therefore largely shared by the two

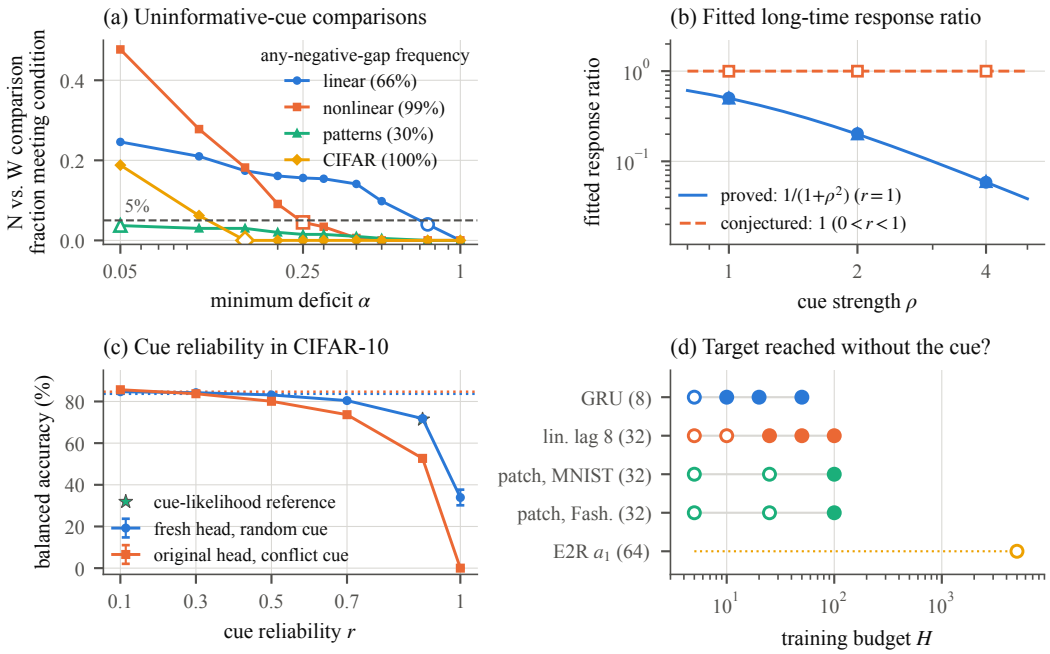


Figure 2: **(a)** Fraction of N-versus-W calibration comparisons meeting the minimum-deficit condition as α increases. Open markers show the selected $\alpha^\star$; the legend gives each family's frequency under the any-negative-gap rule. **(b)** Fitted long-time response ratios, compared with the proved reliable-cue limit and the conjectured imperfect-cue limit. Triangles are theory checks; circles and squares are long runs, summarized by medians over initializations. **(c)** CIFAR accuracy (8 seeds, 95% t-intervals). Dotted lines show W0; the star is a model-based reference combining W0 predictions with cue likelihoods. **(d)** Whether the cue-free run reaches its target at each budget H (learning rate times updates). Open markers mean none reaches it; filled markers mean all do. Counts are in parentheses.

runs under this protocol. The comparison remains inconclusive about suppression; it proves neither intrinsic unlearnability nor an absence of cue effects. At the selected larger amplitudes, the cue-free model reaches the target in 32/32 runs per dataset and B meets the full criterion.

For the noiseless zero-lag linear model, 96/96 fitted limits at perfect cue reliability are within 0.03 of $1/(1+\rho^2)$. With reliability 0.9, the fitted limits are 0.997–0.999, consistent with the conjectured limit of one (Figure 2b). Fitted limits support the conjecture but do not prove it.

CIFAR shows a related finite-time difference between perfect and imperfect cues (Figure 3a). With weight decay removed, training R90 four times longer changes its fresh-head cue-free accuracy gap to W0 from -12.5 to -9.0 points. R100's gap instead changes from -40.1 to -42.0 points. The criterion is met at both budgets. These runs use four seeds and stretch the cosine learning-rate schedule with training duration, so they compare two training protocols rather than establish an infinite-time limit. R90 fits 0.61 of its cue-inconsistent training examples at the standard budget and 0.98 at four times that budget; some improvement may reflect memorization.

5.3 Removing or Replacing the Cue

Changing the cue during training provides a more direct intervention. Starting from the same checkpoint and optimizer state, we continue training with the cue retained, removed, or made independent of the label. After a switch at 25% of training and 471 further updates, cue removal improves fresh-head cue-free accuracy by 7.6 [6.6, 8.7] points relative to the unchanged branch. Randomization gives 7.6 [6.5, 8.8] points (Figure 3b). These estimates are similar in magnitude; we did not test statistical equivalence.

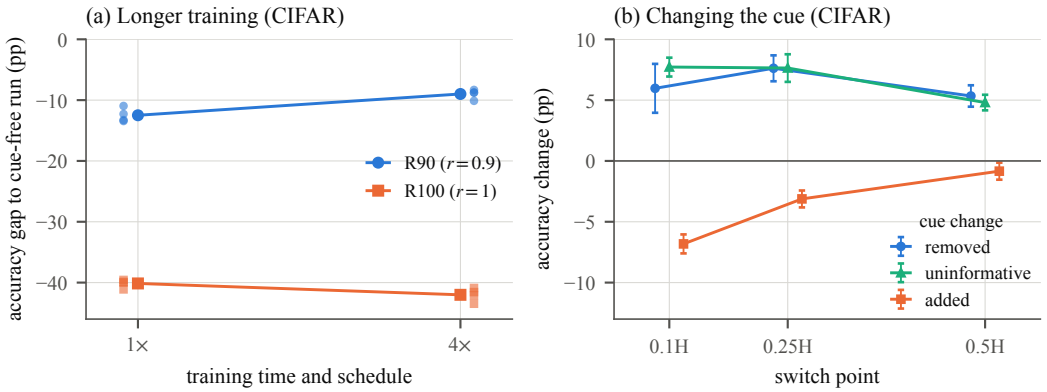


Figure 3: **(a)** CIFAR fresh-head cue-free accuracy gap to W0 at the two training budgets, with the learning-rate schedule stretched accordingly. Dots show individual seeds and lines show means (4 seeds, no weight decay). **(b)** Fresh-head cue-free accuracy change after 471 updates following a cue change, relative to the corresponding unchanged branch (8 seeds, 95% t-intervals). Removing a cue and replacing it with an uninformative one both partly restore accuracy.

Adding an informative cue to a cue-free run changes accuracy by -6.8, -3.1, and -0.8 points at the three switch times. Together, the interventions show that the measured deficit can be partly reversed and that introducing the cue can create a deficit during the remaining training time.

5.4 Feature learning and robustness

Changing the prediction head helps distinguish the features a model has learned from the way it uses them. In CIFAR, R90's cue-free accuracy gap to W0 is -19.4 points with its original head and -11.1 points with a fresh head. Refitting the head reduces the gap but does not remove it. This suggests differences in both the original classifier and the features accessible to a new linear classifier; the two measurements do not form an additive causal decomposition. The fresh-head random-code accuracy gap to R10 reaches -12.87 $[-13.58, -12.16]$ points at cue reliability 0.9 (Figure 2c).

Waterbirds gives a useful contrast (Sagawa et al., 2020). We fine-tune an ImageNet-pretrained ResNet-50 with the official backgrounds, backgrounds independent of the bird label, or gray backgrounds. Bird pixels are identical across these conditions. The cue-free run reaches its target in 8/8 seeds, but none of the informative-background runs meets the full criterion. Fresh-head gray-background accuracy differences are small and nonzero: -0.5 $[-0.8, -0.3]$ points relative to independent-background training and -1.3 relative to gray-background training. Worst-group accuracy nevertheless differs substantially, from 0.917 with independent backgrounds to 0.750 with official backgrounds.

This result is consistent with a model retaining useful core features while relying on background information for prediction, as observed in last-layer retraining studies (Kirichenko et al., 2023). It concerns features available after pretraining and measured during fine-tuning. It does not establish that representations are identical, or that other response measures would give the same conclusion. Additional probe comparisons, architectural replications, and mechanism studies are retained in Appendix E.

6 Discussion and conclusion

The central question is comparative: how much does a cue change what the same training procedure learns about another feature? The cue-free run checks whether that procedure makes enough progress for the comparison to be informative. The uninformative-cue run checks whether the difference depends on label information. Requiring a minimum deficit at the target and over training prevents every small negative difference from counting as suppression.

These choices also limit the interpretation. Results depend on the response, target, thresholds, and budget. Four CIFAR response measures agree in only 5/8 seeds. The pooled control frequencies do

not establish population error rates: seeds are reused across conditions, and several image comparisons have only four or eight seeds. The mathematical results apply under their stated model and smoothness assumptions. The rigorous numerical bounds concern two runs in small networks, and Waterbirds concerns a pretrained backbone. Failed and inconclusive prespecified analyses remain in Appendix F.

The useful distinction is between a feature that the available training procedure learns poorly and a feature that it learns less well because an informative cue is present. Our comparisons make that distinction explicit and show when the chosen evidence is insufficient. They also separate feature learning from the way a trained model uses its features, helping explain why a large robustness failure need not imply a large weak-feature learning deficit.

Reproducibility statement

Appendix B provides proofs under the stated assumptions, and Appendix C lists the numerical checks with their tolerances. Appendix D records the architectures, hyperparameters, seed ranges, calibration rules and deviations, including the units of analysis and limitations of pooled intervals. Appendix E reports supporting experiments and their limitations. An anonymous code repository is linked in the abstract. The curves in Figure 1 are illustrative, not measurements.

AI use statement

We used OpenAI Codex and Anthropic Claude to assist with drafting and polishing manuscript sections, retrieving and identifying related work, exploring research ideas and carrying out experiments, and developing mathematical arguments and proofs. Figure 1 was reconstructed in Adobe Photoshop using editable vector shapes and text, with an AI-generated concept as the visual reference, and checked against Definition 1. Every reported number is read from result files, and every quantitative figure is regenerated from them by a script. The authors take responsibility for the proofs, code, citations and reported results.

References

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson K. Fletcher. Implicit bias of linear RNNs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 2982–2992, 2021. URL https://proceedings.mlr.press/v139/emami21b.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. URL https://doi.org/10.1214/aop/1176996452.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. URL https://doi.org/10.1038/s42256-020-00257-z.

Katherine L. Hermann and Andrew K. Lampinen. What shapes feature representations? Exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9995–10006, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/71e9c6620d381d60196ebe694840aaaa-Abstract.html.

Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bf215fa7fe70a38c5e967e59c44a99d0-Abstract-Conference.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*,

volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.02032.

Fredrik Johansson. Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017. URL https://doi.org/10.1109/TC.2017.2690633.

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://openreview.net/forum?id=Zb6c8A-Fghk.

Alex Krizhevsky. 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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL https://doi.org/10.1109/5.726791.

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*, volume 139 of *PMLR*, pp. 6781–6792, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Rudolf J. Lohner. Enclosing the solutions of ordinary initial and boundary value problems. In Edgar Kaucher, Ulrich Kulisch, and Christian Ullrich (eds.), *Computerarithmetic: Scientific Computation and Programming Languages*, pp. 255–286. B. G. Teubner, Stuttgart, 1987. ISBN 3-519-02448-9. URL https://d-nb.info/870368273/04.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddc3427c5d77843c2253f1e799fe933-Abstract.html.

Nedialko S. Nedialkov, Kenneth R. Jackson, and George F. Corliss. Validated solutions of initial value problems for ordinary differential equations. *Applied Mathematics and Computation*, 105(1):21–68, 1999. URL https://doi.org/10.1016/S0096-3003(98)10083-8.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Aahlad Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! Shortcut learning due to gradients and cross entropy. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71874–71910, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e35460304fdf6df523f068a59aaf8829-Abstract-Conference.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will DNNs choose? A study from the parameter-space perspective. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=qRDQi3ocgR3.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9573–9585, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://www.jmlr.org/papers/v19/18-188.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. URL https://doi.org/10.1109/TPAMI.2017.2723009.

A MEASUREMENTS AND IMPLEMENTATION

Measurements. Synthetic models use the output contrast $M_a = [f(P_a^+) - f(P_a^-)]/2$. The two probe inputs differ only in the sign of a unit feature at the specified time and channel; all other entries are zero. For linear models the contrast is $c^\top W^\ell B e_a$. In the injected-pattern experiments, we average $[f(z + aP_w) - f(z - aP_w)]/2$ over 256 cue-free background images z. Here a is the pattern amplitude and P_w is the weak-feature pattern.

CIFAR uses the original head's class-balanced Brier skill on images with the cue removed:

$$M = 1 - \frac{\frac{1}{10}\sum_{k=1}^{10} \mathbb{E}[\|p - e_k\|^2 \mid y = k]}{1 - \frac{1}{10}}.$$

Here p is the vector of predicted class probabilities and e_k is the one-hot label. Waterbirds uses the corresponding class-balanced score for its two classes.

Evaluation at logged times. Training logs are aligned at $0 = t_0 < t_1 < \cdots < t_K \le H$, with differences $\Delta_k = M_k^X - M_k^Y$. For the any-negative-gap comparison, a difference counts as negative only if $\Delta_k < -\tau_{\text{num}}$, where $\tau_{\text{num}} = 10^{-9}$. The reference first reaches its target at the smallest logged index $k^\star$ with $M_{k^\star}^Y \ge \beta$, provided $\beta > M_0^Y$. The at-target deficit requirement is $\Delta_{k^\star} \le -\epsilon M_{k^\star}^Y$ with $\epsilon = 0.05$. The normalized accumulated deficit uses the trapezoidal rule:

$$a_{X|Y} = \frac{\text{trapz}(M^Y - M^X, t)}{\text{trapz}(\max(M^Y, 0), t)}.$$

A nonpositive denominator gives an undefined area and no positive suppression result. All reported targets are positive.

These calculations approximate Definition 1. A logged crossing can overshoot β, whereas a continuous response satisfies $M^Y(\tau_Y) = \beta$ at the first crossing. The area also depends on the logging grid. Decisions based on logs do not provide the continuous-time guarantees obtained through interval integration.

Original result-file identifiers. The saved data and frozen plans retain their original names. The following mapping connects those names to the terminology used here:

- `nondeg`: $\beta > M_0^Y$; `gate`: this condition holds and the reference reaches its target.
- `mat_hit`: the reference reaches the target and the at-target deficit requirement holds.
- `MATERIAL_CAL`: $\text{MS}_{\alpha^\star}(X \mid Y)$; `MATERIAL`: the first replication's fixed-threshold condition $\text{MS}_{0.10}(X \mid Y)$.
- `CERT3`: the full criterion, $\text{MS}_{\alpha^\star}(\text{B} \mid \text{W}) \wedge \text{MS}_{\alpha^\star}(\text{B} \mid \text{N})$.
- `BROAD`: the reference reaches its target and some $k \ge 1$ has $\Delta_k < -\tau_{\text{num}}$.
- `AT_HIT`: the reference reaches its target and $\Delta_{k^\star} < -\tau_{\text{num}}$.
- `UNGATED_BROAD`: some $k \ge 1$ has $\Delta_k < -\tau_{\text{num}}$, without requiring a target crossing.
- `xing_pos_neg`: some $k_1 < k_2$ have $\Delta_{k_1} > \tau_{\text{num}}$ and $\Delta_{k_2} < -\tau_{\text{num}}$, meaning transfer followed by suppression.

If W does not reach its target, the result is inconclusive because reference progress is insufficient. Otherwise, the stored full-criterion indicator records whether both comparisons meet the requirements.

Threshold grid and calibration data. The candidate values of α are 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.75, and 1.00. If no grid value achieves the calibration target, that family fails the target. F-lin and F-nl pool earlier synthetic runs with a dedicated calibration block; F-e2r pools earlier injected-pattern runs with a calibration block. F-cifar uses R10-versus-W0 records at cue amplitudes 0.5 and 1.0. These runs are disjoint from the confirmation seeds.

Integrated contributions. The rates SW, G, and d are evaluated with one Jacobian–vector product per run over the training set, accumulated in float64. The identity $\text{SW} + \text{G}(+\text{WD}) = d$ is checked

at each attributed state with relative tolerance 10^{-9} in float64 and 10^{-4} in float32. We report $\phi_{\mathrm{SW}} = \mathcal{S}(H)/(\mathcal{S}(H) + \mathcal{G}(H))$, where $\mathcal{S} = \int \mathrm{SW}$ and $\mathcal{G} = \int \mathrm{G}$, when the denominator is nonzero. The terms can have opposite signs; the ratio is an algebraic contribution, not a causal fraction. Weight decay is accounted for separately.

B Mathematical statements and proofs

We use the response M and training comparisons defined in Section 3. Unless stated otherwise, the training loss is

$$L_X(\theta) = \tfrac{1}{n}\sum_{i=1}^{n} w_i \,\mathrm{CE}\big(z_i^X(\theta), y_i\big), \qquad z_i^X(\theta) = f(x_i^X;\theta) \in \mathbb{R}^{d_z}. \tag{4}$$

Here $w_i \geq 0$ are fixed weights of mean 1. The loss is multiclass cross-entropy, or scalar logistic loss when $d_z = 1$ and $y_i \in \{-1, 1\}$. With weight decay $\lambda \geq 0$, gradient flow is $\dot\theta_X = -\nabla L_X(\theta_X) - \lambda\theta_X$. Results involving derivatives require the stated smoothness of f and M.

B.1 Changes in weak-feature response

Let $\sigma(u) = (1 + e^{-u})^{-1}$. Write $\varepsilon_i^X = \partial\,\mathrm{CE}/\partial z\,(z_i^X, y_i)$ for the residual of example i in run X ($\mathrm{softmax}(z) - e_y$ for cross-entropy; $-\sigma(-yf)\,y$ for binary logistic loss), and $g_i^X = J_i^X \nabla M(\theta_X) \in \mathbb{R}^{d_z}$ for its logit sensitivity along the weak response, where $J_i^X = \partial z_i^X/\partial\theta$.

Proposition 1 (Exact response dynamics and drift decomposition)**.** *Under the flow of Equation* (4)*, (i)* $\dot M^X = -\frac{1}{n}\sum_i w_i\langle g_i^X, \varepsilon_i^X\rangle - \lambda\langle\nabla M(\theta_X), \theta_X\rangle$ *at every width and at every state where* f *and* M *are differentiable. (ii) For any pair,* $d_{X|Y} = \mathrm{SW} + \mathrm{G} + \mathrm{WD}$ *exactly, with* $\bar g_i = \frac{1}{2}(g_i^X + g_i^Y)$, $\bar\varepsilon_i = \frac{1}{2}(\varepsilon_i^X + \varepsilon_i^Y)$ *and*

$$\mathrm{SW} = -\tfrac{1}{n}\textstyle\sum_i w_i\langle\bar g_i, \varepsilon_i^X - \varepsilon_i^Y\rangle, \qquad \mathrm{G} = -\tfrac{1}{n}\textstyle\sum_i w_i\langle g_i^X - g_i^Y, \bar\varepsilon_i\rangle,$$

and $\mathrm{WD} = -\lambda\big[\langle\nabla M(\theta_X), \theta_X\rangle - \langle\nabla M(\theta_Y), \theta_Y\rangle\big]$*. (iii) For binary logistic loss with* $w \equiv 1$ *and* $\lambda = 0$*, (i) reads* $\dot M^X = \frac{1}{n}\sum_i \sigma(-y_i f(x_i^X))\,\langle\nabla\big(y_i f(x_i^X)\big), \nabla M\rangle$*: each example pushes the weak response along its margin gradient with weight* $\sigma(-y_i f)$*. (iv) For any update* $\theta \mapsto \theta + u$ *(SGD, momentum, Adam),* $M(\theta + u) - M(\theta) = \langle\nabla M(\theta), u\rangle + R$ *with* $|R| \leq \frac{1}{2}\|u\|^2 \sup_{s\in[0,1]}\|\nabla^2 M(\theta + su)\|$ *whenever* M *is* C^2 *on the segment.*

Proof. (i) By the chain rule, $\dot M^X = \langle\nabla M(\theta_X), \dot\theta_X\rangle$ and $-\nabla L_X = -\frac{1}{n}\sum_i w_i J_i^{X\top}\varepsilon_i^X$, so $\langle\nabla M, -\nabla L_X\rangle = -\frac{1}{n}\sum_i w_i\langle J_i^X\nabla M, \varepsilon_i^X\rangle = -\frac{1}{n}\sum_i w_i\langle g_i^X, \varepsilon_i^X\rangle$. The weight-decay term contributes $-\lambda\langle\nabla M, \theta_X\rangle$. (ii) Subtract (i) for Y from (i) for X and use the identity $g^X\varepsilon^X - g^Y\varepsilon^Y = \bar g(\varepsilon^X - \varepsilon^Y) + (g^X - g^Y)\bar\varepsilon$, which expands to $\frac{1}{2}[2g^X\varepsilon^X - 2g^Y\varepsilon^Y]$. It is applied componentwise in $\mathbb{R}^{d_z}$ and summed with weights w_i. (iii) For logistic loss $\log(1 + e^{-yz})$, the residual is $\varepsilon_i = -\sigma(-y_i f_i)\,y_i$ with $f_i = f(x_i)$, so $-g_i\varepsilon_i = \sigma(-y_i f_i)\langle y_i\nabla f_i, \nabla M\rangle = \sigma(-y_i f_i)\langle\nabla(y_i f_i), \nabla M\rangle$. (iv) Taylor's theorem with the integral remainder along the segment $\theta + su$, $s \in [0, 1]$. □

B.2 Drift balance under stochastic training

Proposition 2 (Paired drift balance under SGD)**.** *Let both runs share their initialization and use a stochastic rule with square-integrable response increments and shared or independent minibatches. Let* $\mathcal{F}_k$ *be the joint history up to step* k*, including optimizer state, and write* $\delta_k = \Delta_{k+1} - \Delta_k = \mu_k + \xi_k$ *with* $\mu_k = \mathbb{E}[\delta_k \mid \mathcal{F}_k]$*. (i)* $\Delta_K = \sum_{k<K}\mu_k + \sum_{k<K}\xi_k = P_K - N_K + \sum_{k<K}\xi_k$*, where* $P_K = \sum[\mu_k]_+$ *and* $N_K = \sum[-\mu_k]_+$*. (ii) For fixed* K*, if* $\xi_k \leq c$ *almost surely for a fixed* $c > 0$ *and* $V_K = \sum_{k<K}\mathbb{E}[\xi_k^2 \mid \mathcal{F}_k]$*, then for* $s, v > 0$*,* $\Pr\big(\sum_{k<K}\xi_k \geq s,\ V_K \leq v\big) \leq \exp\big(-s^2/(2(v + cs/3))\big)$*. Hence, for fixed* $v > 0$ *and* $\delta \in (0, 1)$*, the event* $V_K \leq v$ *and* $\Delta_K > \sum_{k<K}\mu_k + \sqrt{2v\log(1/\delta)} + \frac{2c}{3}\log(1/\delta)$ *has probability at most* δ*. This is a joint-event bound, not conditional coverage given* $V_K \leq v$*. (iii) Writing* ξ_k^X *for run* X*'s centered one-step response increment,* $\mathrm{Var}(\xi_k^{\mathrm{B}} - \xi_k^{\mathrm{W}} \mid \mathcal{F}_k) = \mathrm{Var}(\xi_k^{\mathrm{B}} \mid \mathcal{F}_k) + \mathrm{Var}(\xi_k^{\mathrm{W}} \mid \mathcal{F}_k) - 2\,\mathrm{Cov}(\xi_k^{\mathrm{B}}, \xi_k^{\mathrm{W}} \mid \mathcal{F}_k)$*. Holding the conditional noise distributions of the two runs fixed, a shared-batch coupling reduces paired variance relative to independent sampling exactly when this covariance is positive.*

Proof. (i) is the Doob decomposition of the increments together with $[x]_+ - [-x]_+ = x$. (ii) is Freedman's inequality (Freedman, 1975) applied to the martingale $\sum \xi_k$. Setting the right-hand side equal to δ gives $s^2 = 2L(v + cs/3)$ with $L = \log(1/\delta)$. Its positive root is $s^\star = \frac{cL}{3} + \sqrt{\frac{c^2L^2}{9} + 2vL} \le \sqrt{2vL} + \frac{2cL}{3}$, and for $s \ge s^\star$ the tail bound is at most δ. (iii) is the variance of a difference. The condition is stated, not assumed. Table 3 checks the band with the realized V_K and $c = \max|\xi_k|$ plugged in; we call this the plug-in band. An observed maximum does not supply the fixed almost-sure bound required by (ii), so its empirical coverage is not guaranteed by that inequality. The continuous-time counterpart is $\Delta(t) = \int_0^t [d]_+ - \int_0^t [-d]_+$. □

In a paired logistic-regression model the plug-in band covers 0.999 of runs and shared minibatches reduce the paired variance (0.00192 against 0.00529). On a 10,000-image CIFAR subset trained with SGD, the plug-in band covers 4/16 runs and the estimated shared-to-independent variance ratio is 0.984 [0.967, 1.005], with an interval including one. There the first-order expected drift omits the finite-step remainder not controlled here for steps crossing ReLU boundaries (Proposition 1(iv)), so the band does not give a rigorous bound for these networks.

B.3 The initial learning rate and the uninformative cue

Consider the linear RNN $h_{k+1} = Wh_k + Bx_k$, $f = c^\top h_T$, $h_0 = 0$, on noiseless two-feature sequences with $w_i \equiv 1$: the weak input is y_i, and the cue is ρy_i or $-\rho y_i$. The cue has amplitude $\rho > 0$ at lag 0, and the weak feature has unit amplitude at lag ℓ. The probe responses are $M_s = c^\top b_s$ and $M_w = c^\top W^\ell b_w$, where b_s, b_w are the columns of B. Let $K = b_s^\top W^\ell b_w = \langle \nabla M_w, \nabla M_s \rangle$ be their geometric coupling and $A = \|\nabla M_w\|^2$. At θ_0 let $s = M_s$ and $m = M_w$, and set $q_c = \sigma(-\rho s - m)$, $q_i = \sigma(\rho s - m)$ and $s_{\mathrm{W}} = \sigma(-m)$.

Theorem 1 (Initial rate gap at cue reliability r)**.** *If the cue agrees with the label on a fraction r of examples, balanced within each class, then under logistic gradient flow*

$$d_0(r) := \dot\Delta(0) = r\, q_c(\rho K + A) + (1 - r)\, q_i(A - \rho K) - s_{\mathrm{W}} A. \tag{5}$$

Shared weight decay cancels from the initial paired gap. At $r = 1$ this is $q_c(\rho K + A) - s_{\mathrm{W}} A$. At $s = m = 0$ it is $d_0(r) = \frac{1}{2}(2r - 1)\rho K$. If $d_0 \ne 0$, then $\Delta(t)$ has the sign of d_0 on some interval $(0, t_1)$.

Corollary 1 (The uninformative-cue comparison)**.** *The N run has $r = \frac{1}{2}$. (i) For $m = 0$, $d_0^{\mathrm{N}} = -\frac{1}{2}\rho K \tanh(\rho s/2)$, with small-$|\rho s|$ approximation $-\rho^2 K s/4$, so $\operatorname{sign}(d_0^{\mathrm{N}}) = -\operatorname{sign}(Ks)$, and in general $|d_0^{\mathrm{N}} + \frac{1}{2}\rho K \tanh(\rho s/2)| \le (\frac{1}{8}\rho|K| + \frac{1}{4}A)|m|$. (ii) d_0^{N} is exactly antisymmetric under $c \mapsto -c$, which maps $(K, s, m) \mapsto (K, -s, -m)$. Hence under any initialization law invariant under $c \mapsto -c$ with $\Pr[d_0^{\mathrm{N}} = 0] = 0$, $\Pr[d_0^{\mathrm{N}} < 0] = \frac{1}{2}$: for half of the initializations the uninformative cue alone makes the early paired gap negative, and in continuous time those runs meet the any-negative-gap rule whenever W reaches the target.*

Proof. With $h_0 = 0$ and noiseless inputs, $f(x) = \sum_k c^\top W^{T-1-k} B x_k = x_{T-1,s}\, c^\top b_s + x_{T-1-\ell,w}\, c^\top W^\ell b_w$. For an example with label y_i, weak input y_i and cue $\zeta_i \rho y_i$ ($\zeta_i = +1$ on a fraction r of the examples), the signed logit is $y_i f(x_i) = \zeta_i \rho M_s + M_w$. With $\mathrm{sp}(x) = \log(1 + e^{-x})$, the B loss is therefore $L_{\mathrm{B}} = r\ \mathrm{sp}(\rho M_s + M_w) + (1 - r)\ \mathrm{sp}(-\rho M_s + M_w)$, and $L_{\mathrm{W}} = \mathrm{sp}(M_w)$. With $\lambda = 0$, gradient flow gives

$$\dot M_w^{\mathrm{B}} = r\, q_c \langle \nabla M_w, \rho \nabla M_s + \nabla M_w \rangle + (1 - r)\, q_i \langle \nabla M_w, -\rho \nabla M_s + \nabla M_w \rangle,$$
$$\dot M_w^{\mathrm{W}} = s_{\mathrm{W}} \|\nabla M_w\|^2.$$

The gradients are $\nabla_c M_s = b_s$, $\nabla_{b_s} M_s = c$, $\nabla_c M_w = W^\ell b_w$, $\nabla_{b_w} M_w = (W^\ell)^\top c$ and $\nabla_W M_w = S_w = \sum_{k=0}^{\ell-1} (W^k)^\top c b_w^\top (W^{\ell-1-k})^\top$, with all other blocks zero. The only overlapping block of ∇M_w and ∇M_s is c, so $\langle \nabla M_w, \nabla M_s \rangle = b_s^\top W^\ell b_w = K$ and $\|\nabla M_w\|^2 = A$. For shared weight decay, add $-\lambda \langle \nabla M_w(\theta_X), \theta_X \rangle$ to each drift. At $t = 0$ the two runs share θ_0, so these terms cancel and subtracting gives Equation (5). At $s = m = 0$, $q_c = q_i = s_{\mathrm{W}} = \frac{1}{2}$ and $d_0 = \frac{1}{2}(2r - 1)\rho K$. Finally, $\Delta(0) = 0$ and $\Delta(t) = d_0 t + o(t)$.

For Corollary 1, set $r = \frac{1}{2}$. Then $d_0^{\mathrm{N}} = \frac{1}{2}\rho K(q_c - q_i) + A[\frac{1}{2}(q_c + q_i) - s_{\mathrm{W}}]$. At $m = 0$, $q_c - q_i = \sigma(-\rho s) - \sigma(\rho s) = -\tanh(\rho s/2)$ and $\frac{1}{2}(q_c + q_i) = \frac{1}{2} = s_{\mathrm{W}}$. For general m, differentiate

in m and use $\sigma' \in (0, \frac{1}{4}]$: $|\partial_m(q_c - q_i)| \le \frac{1}{4}$ and $|\partial_m[\frac{1}{2}(q_c + q_i) - s_{\mathrm{W}}]| \le \frac{1}{4}$. The mean value theorem gives the bound $(\frac{1}{8}\rho|K| + \frac{1}{4}A)|m|$. Under $c \mapsto -c$, $s = c^\top b_s$ and $m = c^\top W^\ell b_w$ change sign, while K and A do not depend on the sign of c. Using $\sigma(-x) = 1 - \sigma(x)$, the sign flip sends $(q_c, q_i, s_{\mathrm{W}})$ to $(1-q_c, 1-q_i, 1-s_{\mathrm{W}})$. Substituting into the exact expression gives $d_0^{\mathrm{N}}(-c) = -d_0^{\mathrm{N}}(c)$. Invariance and $\Pr[d_0^{\mathrm{N}} = 0] = 0$ therefore imply $\Pr[d_0^{\mathrm{N}} < 0] = \Pr[d_0^{\mathrm{N}} > 0] = 1/2$. This symmetry is unconditional; conditioning on W reaching the target need not preserve it. □

B.4 Adding the cue at the output

Proposition 3 (Late fusion). *Let the cue bypass all shared parameters, $f_L(x) = f(\tilde{x}; \theta) + u\, x_s$ with $\tilde{x}$ the cue-free input and $u(0) = 0$, and let M be evaluated on cue-free probes. (i) At a differentiable initial state, for any architecture, data and differentiable loss, $d_{\mathrm{B|W}}(0) = 0$ and $\mathrm{SW}(0) = \mathrm{G}(0) = 0$. (ii) For a core with locally Lipschitz gradient, under unregularized logistic gradient flow, if the data are noiseless and rank-one, B's cue is ρy_i ($r = 1$), the weak probe equals the training weak input, and $f(\cdot; \theta)$ is odd in its input (for example a bias-free linear RNN), then both runs follow time-reparametrizations of one flow, $\theta_X(t) = \psi(T_X(t))$ with $\psi' = \nabla M_w(\psi)$, and $T_{\mathrm{B}}(t) \le T_{\mathrm{W}}(t)$. Hence $\Delta_{\mathrm{B|W}}(t) \le 0$ for all t. The cue can only slow weak learning and can never transfer to it.*

Proof. (i) At $t = 0$, $u = 0$, so the two runs have identical parameters and identical logits $f(\tilde{x}_i; \theta_0)$, hence identical residuals. M is evaluated on cue-free probes, so $\partial M/\partial u = 0$. The logit Jacobian in θ is the same function of θ in both runs because $\tilde{x}_i$ is shared. Hence $g_i^{\mathrm{B}} = g_i^{\mathrm{W}}$ and $\varepsilon_i^{\mathrm{B}} = \varepsilon_i^{\mathrm{W}}$, and Proposition 1 gives $\mathrm{SW}(0) = \mathrm{G}(0) = \mathrm{WD}(0) = 0$.

(ii) Every signed training example equals P^+ (the unit weak input at lag ℓ), plus B's fused term ρu. Oddness gives $y_i f(y_i P^+) = f(P^+) = M_w(\theta)$, so B's single signed logit is $M_w(\theta) + \rho u$ and W's is $M_w(\theta)$. Gradient flow gives $\dot{\theta}_{\mathrm{B}} = \sigma(-M_w(\theta_{\mathrm{B}}) - \rho u)\nabla M_w(\theta_{\mathrm{B}})$ with $\dot{u} = \rho\, \sigma(-M_w(\theta_{\mathrm{B}}) - \rho u) > 0$, and $\dot{\theta}_{\mathrm{W}} = \sigma(-M_w(\theta_{\mathrm{W}}))\nabla M_w(\theta_{\mathrm{W}})$ with $u \equiv 0$. Let ψ solve $\psi' = \nabla M_w(\psi)$, $\psi(0) = \theta_0$, on its maximal interval $[0, \tau^\star)$. Define T_{W} by $\dot{T}_{\mathrm{W}} = F(T_{\mathrm{W}})$ with $F(T) = \sigma(-M_w(\psi(T)))$, and define (T_{B}, u) by $\dot{T}_{\mathrm{B}} = \sigma(-M_w(\psi(T_{\mathrm{B}})) - \rho u)$, $\dot{u} = \rho \dot{T}_{\mathrm{B}}$, all starting from 0. Then $\psi(T_X(t))$ solves run X's equation with the same initial value, so by uniqueness $\theta_X(t) = \psi(T_X(t))$. In addition, $T_X(t) < \tau^\star$ for all t. Otherwise θ_X would blow up in finite time, which the finite path length $\int_0^t \|\dot{\theta}\| \le \sqrt{t\, L_X(\theta_0)}$ of gradient flow excludes. Since $u \ge 0$ and $x \mapsto \sigma(-x)$ is decreasing, $\dot{T}_{\mathrm{B}} \le F(T_{\mathrm{B}})$, and F is locally Lipschitz. The comparison lemma therefore gives $T_{\mathrm{B}}(t) \le T_{\mathrm{W}}(t)$. Finally, $\frac{d}{dT} M_w(\psi(T)) = \|\nabla M_w(\psi(T))\|^2 \ge 0$, so $\Delta(t) = M_w(\psi(T_{\mathrm{B}}(t))) - M_w(\psi(T_{\mathrm{W}}(t))) \le 0$. For $t > 0$, $u > 0$ makes the comparison strict, so the inequality is strict whenever $\nabla M_w \ne 0$ along the path. □

B.5 Long-time behavior

For unweighted logistic loss at lag 0 with no background noise, the recurrent matrix receives no gradient (Lemma 1), and the model reduces to the two-layer linear network $f = c^\top B x$ on the span of the strong and weak coordinates. The signed inputs $y_i x_i$ are $\chi_+ = (\rho, 1)$ with weight r and $\chi_- = (-\rho, 1)$ with weight $1 - r$ in B, and $(0, 1)$ in W. Write $\kappa^2 = 1 + \rho^2$. In the single-pattern case (a), $m_X(t)$ denotes run X's training margin. Here $\tau_{\mathrm{W}}(\beta) = \inf\{t \ge 0 : M^{\mathrm{W}}(t) \ge \beta\}$ is the untruncated hitting time; it gives the target-crossing time in Definition 1 whenever $\tau_{\mathrm{W}}(\beta) \le H$.

Theorem 2 (Long-time responses). *Consider logistic gradient flow. In (a) and (b), assume the W growing mode is nonzero, $c_0 + B_0 e_w \ne 0$; in (a), also assume $c_0 + B_0 \chi_+/\kappa \ne 0$, where $e_w = (0, 1)$. (a)* Reliable cue ($r = 1$), no weight decay. *$m_X(t) = \log t + \log\log t + \log(2\kappa_X) + o(1)$ with $\kappa_{\mathrm{B}} = \kappa$ and $\kappa_{\mathrm{W}} = 1$. In addition $M_w^{\mathrm{W}} = m_{\mathrm{W}}$, and the identity $M_w^{\mathrm{B}} = m_{\mathrm{B}}/\kappa^2 + \langle c, e_0 \rangle$ holds with a fixed e_0 and $\|c\| = O(\sqrt{\log t})$. Consequently $M_w^{\mathrm{B}}/M_w^{\mathrm{W}} \to 1/(1 + \rho^2)$, $\Delta(t) = -\frac{\rho^2}{1+\rho^2} \log t\, (1 + o(1)) \to -\infty$ and $a_{\mathrm{B|W}}(H) \to \rho^2/(1 + \rho^2)$. W eventually reaches every target $\beta > \max\{0, M(\theta_0)\}$, with $\log \tau_{\mathrm{W}}(\beta) = \beta - \log\beta - \log 2 + o(1)$ as $\beta \to \infty$. In particular, for any $0 < \epsilon, \alpha < \rho^2/(1 + \rho^2)$ there are β_0, H_0 such that $\mathrm{MS}_{\epsilon,\alpha}(\mathrm{B} \mid \mathrm{W})$ holds whenever $\beta \ge \beta_0$ and $H \ge \max\{\tau_{\mathrm{W}}(\beta), H_0\}$: the B-versus-W suppression condition is* permanent *under these assumptions. (b)* Unreliable cue (empirical fraction $0 < r < 1$), no weight decay. *If the risk falls below $\log 2$ at some time, the end-to-end predictor satisfies $B^\top c/\|B^\top c\| \to (0, 1)$ (Ji & Telgarsky, 2019), so $M_s^{\mathrm{B}}/M_w^{\mathrm{B}} \to 0$. (c)*

Regularized or non-separable models that are linear in parameters. *If the convex risk $L_X + \frac{\lambda}{2}\|\theta\|^2$ has bounded sublevel sets and a unique minimizer $\theta_X^\star$ for each run, then $\Delta(t) \to M(\theta_{\mathrm{B}}^\star) - M(\theta_{\mathrm{W}}^\star)$. If $M(\theta_{\mathrm{W}}^\star) > \beta$, W reaches the target at a finite time. If $M(\theta_{\mathrm{W}}^\star) < \beta$, the weak-only response is eventually below β. This does not exclude an earlier target crossing.*

Conjecture 1 (Transience)**.** *Under the initialization and risk assumptions of case (b), $M_w^{\mathrm{B}}/M_w^{\mathrm{W}} \to 1$ and $a_{\mathrm{B|W}}(H) \to 0$. The B|W minimum-deficit condition, and hence the full suppression criterion, would then eventually fail at any fixed $\alpha > 0$.*

Lemma 1 (Zero-lag closure)**.** *If $\ell = 0$ and the inputs vanish before the last step, then $h_{T-1} \equiv 0$ and $f = c^\top B x_{T-1}$ for every θ. Hence $\nabla_W L \equiv 0$ and, without weight decay, $W(t) = W(0)$; $M_s = c^\top b_s$ and $M_w = c^\top b_w$.*

Proof. $h_0 = 0$ and $x_k = 0$ for $k < T-1$ give $h_{T-1} = 0$, so f does not depend on W. □

Proof of Theorem 2. *(a)* Take a run with a single signed input χ, $\kappa = \|\chi\|$ and margin $m = c^\top B\chi$. Only $v = B\chi$ enters the loss $\mathrm{sp}(m)$. The flow gives $\dot c = \sigma(-m)v$ and $\dot B = \sigma(-m)c\chi^\top$, so $\dot v = \kappa^2\sigma(-m)c$. With $p = c + v/\kappa$ and $q = c - v/\kappa$ we get $\dot p = \kappa\sigma(-m)p$ and $\dot q = -\kappa\sigma(-m)q$. Hence $p = e^{\mathcal{A}}p_0$ and $q = e^{-\mathcal{A}}q_0$ with $\mathcal{A}(t) = \kappa\int_0^t \sigma(-m)$, and

$$m = \tfrac{\kappa}{4}\big(\|p\|^2 - \|q\|^2\big) = \tfrac{\kappa}{4}\big(Pe^{2\mathcal{A}} - Qe^{-2\mathcal{A}}\big), \qquad P = \|p_0\|^2 > 0,\ Q = \|q_0\|^2.$$

If $\mathcal{A}$ were bounded, m would be bounded, so $\dot{\mathcal{A}}$ would be bounded below by a positive constant, a contradiction. Hence $\mathcal{A} \to \infty$ and $m \to \infty$. Differentiating, $\dot m = 2\kappa\sigma(-m)(m + \varepsilon_t)$ with $\varepsilon_t = \frac{\kappa}{2}Qe^{-2\mathcal{A}} \to 0$. For sufficiently large t, $m(t) > 1$. On this range define $\Phi(m) = \int_1^m \frac{dx}{2\kappa x\sigma(-x)}$. Then $\frac{d}{dt}\Phi(m(t)) = 1 + \varepsilon_t/m \to 1$, so $\Phi(m(t)) = t(1 + o(1))$. Since $\Phi(m) = \frac{e^m}{2\kappa m}(1 + O(1/m))$, this gives $m - \log m = \log(2\kappa t) + o(1)$, that is, $m = \log t + \log\log t + \log(2\kappa) + o(1)$.

For the responses, $\dot B = \sigma(-m)c\chi^\top$ gives $B(t) = B_0 + C(t)\chi^\top$ with $C = \int \sigma(-m)c$. Then $M_w = c^\top Be_w = c^\top B_0 e_w + c^\top C\,\chi_w$ and $m = c^\top B_0\chi + c^\top C\,\kappa^2$. With $\chi_w = 1$, eliminating $c^\top C$ gives $M_w = m/\kappa^2 + \langle c, e_0\rangle$ with $e_0 = B_0(e_w - \chi/\kappa^2)$. For W, $\chi = e_w$, so $e_0 = 0$ and $M_w^{\mathrm{W}} = m_{\mathrm{W}}$. Since $\|c\| \le \frac{1}{2}(\|p\| + \|q\|)$ and $e^{2\mathcal{A}} = O(m)$, we have $\|c\| = O(\sqrt{m}) = O(\sqrt{\log t})$. The response ratio, difference, and target-crossing statements follow from $m_{\mathrm{B}}/m_{\mathrm{W}} \to 1$ and $m_{\mathrm{B}} - m_{\mathrm{W}} \to \log\kappa$. For the area, the numerator and denominator of $a(H)$ diverge, and the ratio of their integrands tends to $1 - \kappa^{-2}$, so L'Hôpital's rule gives $a(H) \to \rho^2/(1+\rho^2)$. For the suppression condition, $\tau_{\mathrm{W}}(\beta) \to \infty$ as $\beta \to \infty$, and $\Delta(\tau_{\mathrm{W}})/M^{\mathrm{W}}(\tau_{\mathrm{W}}) = M^{\mathrm{B}}/M^{\mathrm{W}} - 1 \to -\rho^2/(1+\rho^2) < -\epsilon$. The finer expansion in the text follows by substituting $m_{\mathrm{B}} = m_{\mathrm{W}} + \log\kappa + o(1)$.

(b) By Lemma 1 the model is the depth-2 linear network $x \mapsto c^\top Bx$ on $\mathbb{R}^2$. The empirical fraction r is rational, so the weighted risk equals an unweighted risk on duplicated points. The rescaling $\chi' = \chi/\kappa$, $c'(u) = \sqrt{\kappa}\,c(u/\kappa)$ and $B'(u) = \sqrt{\kappa}\,B(u/\kappa)$ preserves the gradient-flow equations and margins while enforcing the cited paper's unit-input-norm convention. Predictor directions are unchanged. The maximum-margin direction of $\{\chi_+, \chi_-\}$ is $\bar u = (0, 1)$ with margin 1, and both points are support vectors spanning $\mathbb{R}^2$ (Assumption 2.7 of Ji & Telgarsky, 2019). Restarting the flow at a time t_0 where the risk is strictly below $\mathrm{sp}(0) = \log 2$ gives their Assumption 1.2: a state with risk below $\log 2$ is not critical, because criticality would force $c = 0$ and hence risk exactly $\log 2$. Their Theorem 2.8 gives $B^\top c/\|B^\top c\| \to \bar u$. That theorem is stated for almost every data set. The genericity is used to obtain their constant $\alpha_{\mathrm{JT}} = \min_{\|\xi\|=1, \xi\perp\bar u}\max_{\chi\in S}\langle\xi, \chi\rangle > 0$ (their Lemma 2.10), which here holds directly: $\alpha_{\mathrm{JT}} = \rho$ before normalization and ρ/κ after it. Then $M_s^{\mathrm{B}}/M_w^{\mathrm{B}} = (B^\top c)_s/(B^\top c)_w \to 0$.

(c) A unique minimizer with bounded sublevel sets makes the gradient-flow trajectory bounded, and every limit point is a critical point, hence $\theta_X^\star$ (a LaSalle argument). So $\theta_X(t) \to \theta_X^\star$. Continuity of M gives the response limits. A limiting W response above β ensures a finite crossing; a limit below β ensures eventual subthreshold behavior, without excluding a crossing before that time. □

Correction at finite times. The identity in case (a) also gives

$$\frac{M_w^{\mathrm{B}}}{M_w^{\mathrm{W}}} = \kappa^{-2}\left(1 + \frac{\log\kappa}{m_{\mathrm{W}}}\right) + O\left(\frac{\|e_0\|}{\sqrt{\log t}}\right) + o(1/\log t),$$

where $\|e_0\|$ is proportional to the initialization scale. The corrections decrease slowly with training time. Table 3 compares this expression with direct integration at $t = 10^9$.

B.6 RIGOROUS BOUNDS OVER SUCCESSIVE TIME INTERVALS

Theorem 3 (Bounds from successive trajectory enclosures). *Let $X = (\theta_{\mathrm{B}}, \theta_{\mathrm{W}})$ solve the paired flow $\dot{X} = V(X)$ with $V \in C^1$, and let $X(0) = (\theta_0, \theta_0)$ for every θ_0 in a box I_0. Suppose $0 = t_0 < \cdots < t_J = H$ and compact sets E_j satisfy $X(t) \in E_j$ for all $t \in [t_{j-1}, t_j]$ and $\theta_0 \in I_0$. Fix $t_a \in [0, H]$ and $\epsilon > 0$. (i) If $\sup_{E_j}\{\Delta + \epsilon M(\theta_{\mathrm{W}})\} \le 0$ for every segment meeting $[t_a, H]$, then $\Delta(t) \le -\epsilon M^{\mathrm{W}}(t)$ on $[t_a, H]$. At times with $M^{\mathrm{W}}(t) > 0$, this is a relative suppression bound. (ii) If $\inf_{E_j} M(\theta_{\mathrm{W}}) \ge \beta > \max\{0, \sup_{I_0} M\}$ for some j, the target is reached with $\tau_{\mathrm{W}} \le t_j$. (iii) If $\inf_{E_j} \dot{\Delta} \ge 0$ on every segment meeting $[t_a, H]$, suppression is excluded there.*

Proof and numerical construction. Each $t \in [t_a, H]$ lies in a segment j that meets $[t_a, H]$, and $X(t) \in E_j$, which gives (i) and (iii). For (ii), $M^{\mathrm{W}}(t) \ge \beta$ on $[t_{j-1}, t_j]$, and $M(\theta_0) < \beta$ places the target strictly above the initial response. The substance is in producing the E_j. A Lohner-type step (Lohner, 1987; Nedialkov et al., 1999) first finds an a-priori box $\tilde{E}$ with $X(t_{j-1}) + [0, h_j]\, V(\tilde{E}) \subseteq \operatorname{int} \tilde{E}$. By the Picard–Lindelöf argument, the solution stays in $\tilde{E}$ on the whole step, so we set $E_j = \tilde{E}$. The endpoint set is then tightened by a mean-value Taylor step with QR-preconditioned wrapping. All arithmetic uses outward-rounded balls (Johansson, 2017). The relation $\theta_{\mathrm{B}}(0) = \theta_{\mathrm{W}}(0)$ is represented by one shared initial box to preserve dependence between the runs and avoid treating their initial uncertainty as independent. The reported implementation checks were to reproduce the construction below, contain a DOP853 reference solution at 100 random times for every configuration, and refuse a false claim. These checks test implementation consistency; rigorous bounds require the validated enclosures themselves.

B.7 A LOCAL BOUND AND ITS TANH-CNN EXAMPLE

Proposition 4 (A bound within one parameter ball). *Let $V = (-\nabla L_{\mathrm{B}}, -\nabla L_{\mathrm{W}})$ be C^1, $d = \mathcal{L}_V \Delta$ and $j = \mathcal{L}_V d$. Suppose a validated enclosure of $X(t_w)$ lies within r_w of $\bar{X}_w$, and on $U_R = \{\|X - \bar{X}_w\| \le R\}$ we have $\|V\| \le S$, $j \le -\Lambda < 0$, $\dot{M}^{\mathrm{W}} \ge \nu > 0$ and $r_w + Sh \le R$ with $h = H - t_w \ge 0$. If $\Delta(t_w) \le D_+$, $d(t_w) \le d_+$ and $M^{\mathrm{W}}(t_w) \ge m$, then $\Delta(t_w + u) \le Q(u) = D_+ + d_+ u - \Lambda u^2/2$ and $M^{\mathrm{W}}(t_w + u) \ge m + \nu u$ for $u \in [0, h]$. If $u_a \in [0, h]$, $Q(u_a) < 0$ and $d_+ - \Lambda u_a \le 0$, suppression persists on $[t_w + u_a, H]$. Conversely, $|j| \le J$, $\Delta(t_w) \ge D_- \ge 0$, $d(t_w) \ge d_-$ and $d_- - Jh > 0$ exclude suppression on $[t_w, H]$.*

Proof. The paired field is locally Lipschitz. Before exiting U_R, $\|X(t_w + u) - \bar{X}_w\| \le r_w + Su \le R$, so no exit occurs before H. On U_R, $d' = j \le -\Lambda$. Integrating twice gives Q, and integrating $\dot{M}^{\mathrm{W}} \ge \nu$ gives the lower bound on reference progress. For $u \ge u_a$, $Q'(u) \le 0$. The converse integrates $d' \ge -J$. □

With nonnegative cutoff bounds $D_+, d_+ \ge 0$, the deficit established by this quadratic bound is at most $\Lambda h^2/2 \le \Lambda R^2/(2S^2)$ for $S > 0$. This limit on the verified deficit motivates Theorem 3.

The tanh-CNN construction (configuration L0). Consider $f_\theta(x_s, x_w) = \frac{c}{2}\sum_{p=0}^{1} \tanh(s_0 x_{s,p} + s_1 x_{s,p+1} + w_0 x_{w,p} + w_1 x_{w,p+1})$ with all five parameters trained. B uses $x_s = (y, y, 0)$ and $x_w = (0, y, y)$, W sets $x_s = 0$, and the probe is the odd contrast at $x_w = \pm(0, 1, 1)$. Oddness gives the signed logits $y f_{\mathrm{B}} = \frac{c}{2}[\tanh(s_0 + s_1 + w_1) + \tanh(s_0 + w_0 + w_1)]$ in B and $y f_{\mathrm{W}} = M = \frac{c}{2}[\tanh(w_1) + \tanh(w_0 + w_1)]$. From $\theta_0 = (-10^{-4}, -5 \cdot 10^{-5}, 0, 0, 1)$, with $H = 1/250$ and $\beta = 3/2000$, an outward-rounded enclosure at $t_w = 10^{-4}$ verifies the premises of Proposition 4 with $R = \frac{1}{200}$, $S \le \frac{5}{4}$, $r_w \le 10^{-6}$, $-\frac{3}{10} \le j \le -\frac{3}{32}$, $\dot{M}^{\mathrm{W}} \ge \frac{1}{2}$, $D_+ \le \frac{3}{5\cdot 10^7}$, $d_+ \le \frac{1}{5\cdot 10^4}$ and $m \ge \frac{3}{5\cdot 10^4}$. With $h = \frac{39}{10^4}$ and $u_a = \frac{29}{10^4}$, exact rational arithmetic gives $r_w + Sh = \frac{1219}{250000} < \frac{1}{200}$, $Q(u_a) \le -\frac{8839}{3.2\cdot 10^{10}} < -10^{-7}$, $Q'(u_a) \le -\frac{403}{1.6\cdot 10^6} < 0$ and $m + \nu h \ge \frac{201}{10^5} > \beta$. A warm-up enclosure ($d(0) \ge \frac{3}{10^5}$, $j \ge -\frac{1}{4}$) gives $\Delta(t) > 0$ on $(0, 10^{-4}]$: transfer, then suppression. The verified deficit is only 10^{-7}, about 10^{-4} of the weak-only progress. This small verified margin motivates the chained construction; Section 4.4 uses Theorem 3.

B.8 Additional bounds and prediction checks

Tail-area criterion. If $d > 0$ on $(0, \tau_d)$ and $d < 0$ on (τ_d, H), then $\Delta(t) = A^+ - \int_{\tau_d}^t [-d]_+$ with $A^+ = \Delta(\tau_d)$, so the response difference becomes negative by H if and only if the tail area exceeds A^+. This follows immediately from $\Delta = \int d$. As a forecaster of the crossing time it does not establish an improvement over linear extrapolation (pre-registered, not met; Appendix E).

Conditional tangent comparison. Consider binary logistic gradient flow with $w_i = 1$ and $\lambda = 0$. Let $m_i = y_i f(x_i; \theta)$, let J_m have rows $\nabla m_i^\top$, and set $\Theta = J_m J_m^\top$ and $c_w = J_m \nabla M$. All vector norms are Euclidean and matrix norms are induced operator norms. Freezing these quantities at initialization gives $\dot{\bar{m}} = \Theta_0 \sigma(-\bar{m})/n$ and $\dot{\bar{M}} = c_{w,0}^\top \sigma(-\bar{m})/n$, with $\bar{m}(0) = m(0)$ and $\bar{M}(0) = M(0)$. If $\|\Theta(u) - \Theta_0\| \le B_\Theta$, $\|c_w(u) - c_{w,0}\| \le B_c$ and $\|\sigma(-m(u))\| \le V$ on $[0, t]$, then

$$|M(t) - \bar{M}(t)| \le \frac{B_c V}{n} t + \frac{\|c_{w,0}\| B_\Theta V}{4n^2} \psi_a(t), \qquad a = \frac{\|\Theta_0\|}{4n},$$

where $\psi_a(t) = (e^{at} - 1 - at)/a^2$ for $a > 0$ and $\psi_0(t) = t^2/2$. To see this, $\sigma' \le 1/4$ and Grönwall's inequality give $\|m(t) - \bar{m}(t)\| \le (B_\Theta V/n)(e^{at} - 1)/a$, interpreted continuously at $a = 0$. Integrating the corresponding response-drift bound gives the display. The tested bounds are vacuous (Table 3).

B.9 Counterfactual drift correction (CDC)

At B's state, let $a = \nabla M_s \neq 0$, $p = \nabla M_w$ and $v = -\nabla L_\mathrm{B}$. Set $q = p - \frac{p \cdot a}{\|a\|^2} a$ and $\delta = [F^\mathrm{W} - p \cdot v]_+$, where F^W is the matched weak-only drift. If $\delta = 0$, define $\gamma = 0$ and $v' = v$. If $\delta > 0$, assume $q \neq 0$ and set $\gamma = \delta / \|q\|^2$, $v' = v + \gamma q$. When $q = 0$ and $\delta > 0$, the weak-progress constraint below is infeasible while preserving $a \cdot v$.

Proposition 5 (CDC local guarantees)**.** *Under the preceding nondegeneracy conditions, (1) $a \cdot v' = a \cdot v$. (2) γq is the unique minimum-norm d with $a \cdot d = 0$ and $p \cdot (v + d) \ge F^\mathrm{W}$. (3) If ∇M_s is L-Lipschitz, then $|M_s(\theta + \eta v') - M_s(\theta + \eta v)| \le \frac{L}{2} \eta^2 (\|v'\|^2 + \|v\|^2)$.*

Proof. (1) $a \cdot q = 0$. (2) If $\delta = 0$, then $d = 0 = \gamma q$. Otherwise the constraint is active. Stationarity of $\frac{1}{2}\|d\|^2 - \lambda_1 a \cdot d - \lambda_2 (p \cdot d - \delta)$ gives $d = \lambda_1 a + \lambda_2 p$. The condition $a \cdot d = 0$ forces $d = \lambda_2 q$, and $p \cdot d = \delta$ gives $\lambda_2 = \delta / \|q\|^2$. Strict convexity gives uniqueness. (3) Expand both sides to second order. The first-order terms cancel by (1). □

These guarantees are instantaneous only. In the pre-registered finite-horizon test, Richardson extrapolation in η estimates a nonzero continuous-time limit of the strong-response gap between CDC and plain training (-0.433 $[-0.491, -0.381]$). CDC also costs more strong-cue response than an unconstrained rescue in every synthetic architecture (Table 6). We therefore do not propose CDC as a remedy. It is a local probe whose local identities agree with numerical checks in 60/60 tested states.

C NUMERICAL VERIFICATION OF THE THEORY

Table 3 lists the pre-registered numerical checks (float64 throughout), followed by additional checks of the revised statements. These tests can expose implementation or algebra errors but do not replace the proofs or validate every assumption of an empirical application. The additional checks are not pre-registered.

Table 2: Formal statements, the quantity checked, and the observation (95% intervals). Identities are compared with autograd and asymptotic results with fitted limits; the lower block lists a conjecture and empirical assessments. All pre-registered endpoints are listed in Appendix F.

Statement / study	Quantity checked	Observation
Prop. 1	exact drift decomposition	rel. err. 1.6×10^{-15}; checked at attributed differentiable states
Thm. 1	initial gap $d_0(r)$ and early sign	rel. err. 6.8×10^{-10}; sign correct in 46,000/46,078 evaluated comparisons
Cor. 1	N|W early-gap sign $-\mathrm{sign}(Ks)$	94/96 records; 365/384 in a replication
Prop. 3	zero initial gap; no transfer under the stated premises	$\max \lvert d_0 \rvert = 0$; transfer then suppression in 0/192 (noisy) runs
Thm. 2(a)	fitted ratio limit $1/(1+\rho^2)$, $r = 1$	96/96 fits within 0.03 (medians 0.503, 0.203, 0.0586)
Thm. 2(c)	regularized limit set by the optimum	limit error 1.9×10^{-7} against the convex solver
Thm. 3	rigorous B-versus-W bounds	verified for all 9 configurations; 41–51% of sampled initializations
Conj. 1	fitted ratio limit 1, $r = 0.9$	medians 0.997–0.999
Noisy-label target comparison	target crossing predicted from the end-to-end convex optimum	160/160 predicted passes agree; grid has no predicted failures
CIFAR budget comparison	gap after 4× longer training (finite horizon)	R100 $-40.1 \to -42.0$ pp; R90 $-12.5 \to -9.0$ pp
Initial-rate forecast	AUROC of $-d_0$ for later suppression outcomes	0.617 [0.604, 0.630]; 0.649 at width 32, 0.579 at width 256
Def. 1	control frequency; conditional full-criterion frequency	5.3–6.9% (any-negative-gap rule: 31–100%); 0.980–1.000 (0.667 linear)

The stated $O(\eta^2)$ bound assumes a C^2 response on the update segment. The fitted exponent meets the criterion with GELU and average pooling and fails with ReLU and max-pooling in this check; a piecewise-smooth model can still satisfy such a bound on a segment that stays inside one smooth region. The target-crossing time law $\log \tau_{\mathrm{W}} = \beta - \log \beta - \log 2 + o(1)$ is asymptotic, and at $\beta \leq 10$ the $\log \beta$ term keeps $\log \tau$ visibly nonlinear in β. Across the pre-registered asymptotic runs (A3a), $\log \tau$ against β has median R^2 0.997 and slope 0.759.

Table 3: Theory verification. T-M checks were pre-registered with the criteria shown. The last block contains additional checks of Corollary 1, Proposition 3, and Theorem 2, which were not pre-registered.

Statement	Check (criterion)	Result	Status
Proposition 1(i)	autograd directional derivative, CNN4-GN, weighted CE, $\lambda > 0$ ($\leq 10^{-10}$)	1.9×10^{-16}	met
Proposition 1(ii)	SW + G + WD $= d$ at 10 random paired states ($\leq 10^{-10}$)	1.6×10^{-15}	met
Proposition 1(iv)	slope of $\log \lvert R \rvert$ on $\log \eta$ in $[1.8, 2.2]$	1.575 ReLU/max-pool; 2.0004 GELU/avg-pool	not met
Proposition 2(ii–iii)	plug-in band coverage ≥ 0.95; shared-batch variance smaller	0.999; 0.00192 vs. 0.00529	met
Theorem 1, $r = \frac{1}{2}$	closed form vs. autograd, 1,000 inits ($\leq 10^{-9}$)	4.7×10^{-11}	met
Corollary 1, $m = 0$	symbolic identity $d_0^{\mathrm{N}} = -\frac{1}{2}\rho K \tanh(\rho s/2)$	exact	met
Corollary 1 sign	sign $\Delta_{\mathrm{N\|W}} = -\mathrm{sign}(Ks)$ at early t (≥ 0.90 per ρ)	98.25, 98.25, 98.5, 98.5%	met
integrated control gap	conjecture $\int d^{\mathrm{N}} \approx -Ks/G_{ss}$ (report)	median ratio ≈ 0	not supported
Theorem 2(a), Conjecture 1	fitted R_∞ within 0.02 of the limit ($r \in \{1, 0.9, 0.75\}$)	288/288; medians 0.501, 0.201, 0.0595	met
Theorem 2(a) target-crossing time	$\log \tau$ vs. $\beta \in \{1..10\}$, $R^2 \geq 0.99$	min R^2 0.954, slope 0.59	not met
Theorem 2(c)	label-noise limits vs. convex solver ($\leq 10^{-3}$)	1.9×10^{-7}	met
Theorem 1, general r	closed form vs. autograd; symbolic identities ($\leq 10^{-9}$)	6.8×10^{-10}	met
Theorem 3	integrator implementation checks D1-T1–T3	all pass	met
Lemma 1	$\lVert \nabla_W L \rVert \leq 10^{-12}$ at zero lag; six-scalar ODE matches GD at widths 8–1024	all pass	met
closure residual	$\lvert R_w \rvert \leq 10^{-10}$ at zero background noise (linear)	all pass	met
tangent bound (App. B.8)	bound below the decision margin, per (architecture, width, scale)	0/21 cells	vacuous
Proposition 1(ii), Theorem 1	symbolic: split identity; $d_0(r)$ at $s{=}m{=}0$; $r = \frac{1}{2}$ tanh form; Freedman root	exact	pass
Corollary 1 bound	$2 \cdot 10^5$ random (ρ, K, A, s, m): deviation/bound ≤ 1	max 0.997	pass
Proposition 3(ii)	noiseless late-fusion linear RNN, 24 runs, GD: $\Delta_k \leq 0$ for all k; $\Delta_1 = 0$	$\max_k \Delta_k = 0$	pass
Proposition 3, beyond premises	same with background noise $\sigma = 0.5$: runs with transfer then suppression	0/12	(consistent)
Theorem 2(a) identity	$\lvert M_w^{\mathrm{B}} - m_{\mathrm{B}}/\kappa^2 - \langle c, e_0 \rangle \rvert$ along DOP853 flow to $t = 10^9$	$\leq 7 \times 10^{-15}$	pass
Theorem 2(a) expansion	ratio at $t = 10^9$ vs. $\kappa^{-2}(1 + \log \kappa / m_{\mathrm{W}})$, $\rho = 1, 2, 4$	0.507/0.207/0.063 vs. 0.507/0.207/0.062	pass
Conjecture 1	$r = 0.9$, ratio at $t = 10^{2.25}, 10^{4.5}, 10^{6.75}, 10^9$	0.92, 0.96, 0.97, 0.98	(consistent)

D EXPERIMENTAL DETAILS

Pre-registration. The first replication (seed blocks 34001–34008 for images, 35001–35008 for synthetic runs, and data seeds 36001–36004 × model seeds 36101–36108 for E2R) ran under a single frozen, hashed analysis plan. The second study froze its plans in two hashed stages. Stage 1 fixed every calibration rule and every experiment that needs no calibrated value. Stage 2 fixed the calibrated values ($\alpha^\star$ per family, response-measure thresholds) and every confirmatory design before any confirmatory seed ran. Calibration and confirmation seeds are disjoint: synthetic 41001–41016 and 42001–42032; E2R data 44901–44902 and 44001–44004; images 40001–40004 and 43001–43008; Waterbirds 45901–45904 and 45001–45008. A crashed run was retried once with the identical seed, and a nonfinite outcome counts as a failed run, not as a decision.

Scope reductions, frozen before the affected seeds ran. Full-batch CIFAR (B1, B4, B7), the loss-family study (B3) and the long-horizon study (H3) use 4 seeds; H3 uses $m \in \{1, 4\}$ only. Waterbirds uses the pretrained regime only. The D2 plan specified 200 initializations on the two largest configurations; we report the completed 1,000-initialization runs (the 200-initialization estimates are 0.465 and 0.405). The full-batch CIFAR variants of the tangent and CDC studies, CelebA, MultiNLI, ViT, SST-2 and ImageNet-100 were not run. The σ-damp and σ-stitch losses of Puli et al. (2023) were not included in the CIFAR loss-family study; we did not implement multiclass extensions of their binary-margin formulations.

Units of analysis and intervals. For A1, the N|W control frequency uses every confirmation condition: 640 F-lin, 320 F-nl, 320 F-e2r and 8 F-cifar records. The frozen A1-P2 full-criterion endpoint uses B records whose paired W run reaches the target with $\rho \le 4$ and $\ell \le 4$ for F-lin, and all conditions with a target crossing in W for the other families. Its denominators are 288, 319, 98 and 8. The synthetic grids reuse 32 model seeds across conditions; E2R crosses 4 data seeds and 8 model seeds across 10 dataset–amplitude conditions. The pooled record-level binomial intervals in Appendix E do not account for these dependencies or heterogeneous cell probabilities, so they are nominal summaries rather than calibrated uncertainty for a new seed or condition. Paired CIFAR contrasts use the stated seed-level t-intervals; the fusion ratio uses a seed bootstrap. No ground-truth set of starved runs is available, so A1-P2 is a conditional frequency, not sensitivity.

Training budgets and targets. For synthetic and injected-pattern runs, time is learning rate times updates, $H = \eta \times$ updates. The standard settings are $(\beta, H) = (0.25, 5)$ for F-lin, $(0.5, 20)$ for F-nl, and $(0.5, 100)$ for F-e2r. CIFAR uses $\beta = 0.5$ and counts time in updates. Studies with other budgets are identified below.

Synthetic families (F-lin, F-nl). $n = 512$ sequences, $T = 10$, input dimension 6. Channel 0 is the cue at lag 0, and channel 1 is the weak feature y_i at lag ℓ. Channels 2–5 carry $\mathcal{N}(0, 0.5^2)$ background at every step. The cue is ρy_i (B), 0 (W), or $\rho\xi_i$ with $\xi_i = \pm 1$ balanced within each class (N). Linear RNN: width 96, $W \sim \mathcal{N}(0, g^2/N)$ with $g = 0.5$, $B, c \sim \mathcal{N}(0, 1/N)$, no bias, $\eta = 0.1$. Tanh RNN and GRU: width 32, $g = 0.8$, $\eta = 0.05$. Training is full-batch gradient descent on the mean logistic loss in float64. The linear cells are $\rho \in \{1, 2, 4, 8, 16\} \times \ell \in \{0, 2, 4, 8\}$, and the nonlinear cells are $(\rho, \ell) \in \{(2, 2), (4, 2), (8, 2), (4, 0), (4, 4)\}$.

E2R (F-e2r). MNIST 3-vs-8 and Fashion-MNIST T-shirt-vs-shirt. For each data seed, 2,000 training images are drawn, with labels assigned independently of the image ("shuffled" cores, so the unmodified pixels have no systematic label signal). A per-seed Rademacher pattern P_w is added with amplitude $a \in \{0.001, 0.002, 0.005, 0.01, 0.02\}$ as $a\, y_i P_w$. The cue overwrites the top-left 4×4 patch with a Hadamard code, $y_i Q$ in B and $\xi_i Q$ in N, and is absent in W. The model is a 784–256–1 GELU MLP trained by full-batch gradient descent with $\eta = 0.1$ in float64. M is the probe contrast averaged over 256 cue-free cores.

The first replication required a pair of pattern amplitudes with different outcomes. At the smaller amplitude a_1, it required B's cue response to be at least 1 and its weak response below $\beta = 0.5$ in at least 75% of records. It also required W to reach the target in at most 10% of records at each $H \in \{100, 250, 500\}$. At $H = 100$, the ratio of the mean paired deficit $M^{\mathrm{W}} - M^{\mathrm{B}}$ to the mean shortfall $\beta - M^{\mathrm{B}}$ had to have an upper bootstrap confidence bound of at most 0.20. At the larger

amplitude a_2, W had to reach the target in at least 90% of records and the B-versus-W minimum-deficit condition had to hold in at least 75%. The same cue-response and weak-response requirements on B applied at both amplitudes. Both datasets met these frozen requirements ($a_1 = 0.001$; $a_2 = 0.02$ for MNIST and 0.01 for Fashion). The patch CNN of Figure 2d is an earlier two-channel CNN study on MNIST and Fashion-MNIST, extended from $H = 5$ to $H = 100$.

CIFAR-10 (F-cifar). Splits from the 50,000 training images: 40,000 train, 5,000 validation, 2,500 probe-fit and 2,500 diagnostic. The 32×32 core (reflect-pad 4, random crop, horizontal flip) gets an appended 8-row ribbon, giving 3×40×32 inputs. The ribbon carries one of ten 4×4 Hadamard codes at amplitude 0.5, at a seed-fixed location. Run Rp assigns the true class's code with probability $r = p/100$ and a uniformly chosen wrong code otherwise, so R10 is label-independent in distribution (run N). W0 has a zero ribbon. The model is CNN4-GN: four conv–GroupNorm(8)–ReLU–maxpool blocks with (32, 64, 128, 128) channels, mean pooling, and a linear head. It is trained with SGD (learning rate 0.05, momentum 0.9, weight decay $5 \cdot 10^{-4}$, batch 256, cosine schedule, 4,710 updates). Evaluation conditions: neutral (zero ribbon), consistent, conflict (a fixed wrong code) and random codes. Fresh heads are zero-initialized softmax heads on frozen features, fit on probe-fit with $\ell_2 \in \{0, 10^{-3}, 10^{-2}\}$ chosen on validation. Model-based cue-likelihood combiner: W0's temperature-scaled class probabilities are multiplied by the cue likelihood at $r = 0.9$ (0.9 for the matching code, $0.1/9$ otherwise), then normalized. This assumes a calibrated core predictor and the specified cue model; it is not a known Bayes optimum. Replications: weight decay 0; ResNet-18 (learning rate 0.1); and MNIST–CIFAR dominoes (automobile vs. truck, with MNIST 0/1 on top, 960 updates).

Late fusion (Appendix E.3). The encoder sees only the 32×32 core, and logits are $\mathrm{head}(\mathrm{features}(\mathrm{core})) + U(0.5\, e_{c_i})$ with $U \in \mathbb{R}^{10\times 10}$ zero-initialized and trained like every other parameter. The fusion ratio is $\rho_L = \Delta\mathrm{fh}(\mathrm{R90_L} - \mathrm{W_L})/\Delta\mathrm{fh}(\mathrm{R90} - \mathrm{W0})$ on final fresh-head neutral balanced accuracy, estimated as a ratio of seed means with a bootstrap interval. The frozen decision rule reads "sample weights carry the majority" if the lower bound exceeds 0.5, "geometry carries the majority" if the upper bound is below 0.5, and "mixed" otherwise. These labels are retained as historical endpoint decisions in Appendix F; the ratio does not identify either causal fraction. In particular, the construction does not keep $\mathrm{G} = 0$ beyond initialization. The synthetic version adds $u\, x_{s,T-1}$ with $u(0) = 0$ and zeroes channel 0 for the core.

Full-batch CIFAR. 1,000 images per class, no augmentation, plain gradient descent with $\eta = 0.02$ for $H = 8{,}000$ steps (pilot rule, frozen at Stage 1). The gradient is accumulated over 40 fixed chunks. Its check that the per-step second-order remainders telescope within 5% of $\max|\Delta|$ fails on every seed (median ratio 2,150). A step crossing a nonsmooth boundary need not satisfy the C^2 remainder bound; this empirical failure is why these full-batch attributions are labelled first-order only.

Asymptotics (A3). A3a uses the zero-lag noiseless linear RNN (width 96) with $\rho \in \{1, 2, 4\}$ and $r \in \{1, 0.9\}$, runs to $t = 10^5$, and fits $\mathrm{ratio}(t) = R_\infty + c/\log t$ on $[10^2, 10^5]$. A3c adds label noise $p \in \{0.1, 0.2\}$ with weak amplitude $a_w \in \{0.05, \ldots, 1\}$ and predicts the target crossing at $t \to \infty$ from the convex optimum of the end-to-end predictor. On this grid the optimum's weak coefficient always exceeds β, so the test can confirm only predicted passes. This is an empirical comparison for the factorized model, not an application of Theorem 2(c). A3d runs E2R for 50,000 steps with weight decay $\lambda \in \{0, 10^{-3}, 10^{-2}\}$. A plateau means $|M^{\mathrm{W}}(5000) - M^{\mathrm{W}}(2500)| \le 0.01$ with $M^{\mathrm{W}}(5000) < 0.5$.

Rate-gap studies (C1–C4). C1 covers the noiseless linear RNN over $\rho \in \{0.5, \ldots, 16\} \times \ell \in \{0, 1, 2, 4, 8\} \times g \in \{0.3, 0.5, 0.8, 1\} \times N \in \{32, 96, 256\}$, with 128 initializations per cell, $\eta = 0.01$ and $H = 5$. C3 takes lags up to 16 ($T = 20$) with gains $\{0.3, 0.5, 0.8, 1\}$ and 16 seeds. 76/576 runs became nonfinite in the W run under full-batch gradient descent, all at lag ≥ 3. They are reported as failed. 27 completed records had $M^{\mathrm{W}} \le 0$ throughout, which makes the normalized area undefined, and these were excluded from cell means (logged deviation). C4 takes reliabilities $r \in \{0.5, 0.6, 0.75, 0.9, 1\}$.

Rigorous B-versus-W bounds (D1, D2). The configurations are: L0, the tanh CNN of Appendix B.7 ($P = 5$); L1, a dense linear RNN of width 3 ($P = 18$); L2–L4, linear RNNs with two

inputs, $T = 4$ and width 2, 4, 6 ($P = 10, 28, 54$); L5, a tanh RNN of width 4 ($P = 32$); L6–L8, two-layer tanh MLPs ($P = 49, 97, 145$). All use noiseless inputs, logistic loss and gradient flow with every parameter trained. For each configuration, $\beta = 0.25$ and H is the numerical time at which W reaches 2β, with $t_a = H/2$. The target is $\sup(\Delta + 0.1M^{\mathrm{W}}) \leq 0$ on $[t_a, H]$, plus the W target crossing. Thus the horizon is selected separately for each initialization, before validating the paired trajectory; the resulting frequency is specific to this rule. The bound applies only on the stated interval and does not establish the full suppression criterion. Initializations are drawn from each configuration's initialization law and rounded to a 2^{-30} dyadic grid. "Undetermined" includes initializations for which W never reached 2β or $M(\theta_0) \geq \beta$.

Loss family and crossed fields. Loss-family runs use the CIFAR setup with cross-entropy, label smoothing ($\epsilon = 0.1, 0.3$), spectral decoupling (cross-entropy plus $\frac{\lambda}{2}$ times the mean squared logit norm, $\lambda = 10^{-3}, 10^{-2}, 10^{-1}$), focal loss ($\gamma = 2$) and MSE on logits, with runs W0, R10 and R90 and 4 seeds. The residual-collapse index is $1 - \int \bar{\varepsilon}^{\mathrm{R90}} / \int \bar{\varepsilon}^{\mathrm{W0}}$, with $\bar{\varepsilon}$ the mean residual norm over the unaugmented training set. Crossed fields fix B's state and continue plain gradient descent for $0.05H$ under the field built from cue-on Jacobians and cue-off residuals (BW) or the reverse (WB); $C = M(\mathrm{BW}) - M(\mathrm{WB})$.

Long horizon and cue switching. Long-horizon runs use CNN4-GN without weight decay and the cosine schedule stretched by $m \in \{1, 4\}$ (30 and 120 epochs), 4 seeds. Cue switching starts from R90 or W0 checkpoints (with optimizer state) at 10%, 25% and 50% of training and continues for 471 updates with the ribbon removed, kept, made label-independent (R10 codes) or added (R90 codes), 8 seeds.

Waterbirds. Following Sagawa et al. (2020), we combine CUB-200-2011 bird images (Wah et al., 2011) with Places backgrounds (Zhou et al., 2018). Each image is composited once with one land and one water background from the official category lists. WB follows the official 95% background–label correlation, WR draws the background type independently of the label, and WG uses a gray background; bird pixels are bitwise identical across runs. The backbone is an ImageNet-pretrained ResNet-50 fine-tuned with SGD (learning rate 10^{-3}, 50 epochs), 8 seeds. M is the original-head class-balanced Brier skill on gray-background test images, with $\beta = 0.40$ and $\alpha^\star = 0.05$ calibrated on four WR|WG records. A reweighting counterfactual (weights that make background and label independent) is trained on WB images with shared minibatches.

Optimizer and normalization. AdamW (learning rate 10^{-3}, weight decay 0.05) and a BatchNorm variant of CNN4 use the standard CIFAR pipeline with runs W0, R10 and R90, 8 seeds.

E Additional results

This appendix reports the supporting comparisons and checks omitted from the main narrative. Experiment identifiers refer to the frozen analysis plans and saved result files.

E.1 Control comparisons and threshold choices

Uncertainty in the confirmation frequencies (A1). Table 4 retains the original record-level binomial intervals. They do not adjust for the repeated seeds described in Appendix D. The outcomes are the N-versus-W minimum-deficit condition and the full B-versus-W-and-N criterion. Neither column estimates the criterion's population false-positive rate or ground-truth sensitivity.

What the N comparison adds. At the frozen thresholds, among the A1-P2 records for which W reaches the target, the B-versus-N condition removes 0/192 positive B-versus-W comparisons in F-lin, 1/319 in F-nl, 2/98 in F-e2r, and 0/8 in F-cifar: 3 of 617 in total. This post-hoc count uses existing confirmation records and involves no new training runs. It separates the additional changes due to N from the effect of imposing minimum deficits.

Other thresholds. With a fixed $\alpha = 0.10$, 0.210 of linear and 0.278 of nonlinear calibration comparisons meet the N-versus-W condition. A 10% calibration target would select thresholds giving

Table 4: Original nominal 95% record-level intervals, retained for transparency. F-lin full-criterion frequency uses the prespecified $\rho \leq 4, \ell \leq 4$ subset; all other families use all conditions where W reaches the target.

Family	N|W comparison fraction (95% interval)	Full criterion, W reaches target fraction (95% interval)
F-lin	0.053 [0.037, 0.073]	0.667 [0.609, 0.721]
F-nl	0.069 [0.044, 0.102]	0.997 [0.983, 1.000]
F-e2r	0.053 [0.031, 0.084]	0.980 [0.928, 0.998]
F-cifar	0.000 [0.000, 0.369]	1.000 [0.631, 1.000]

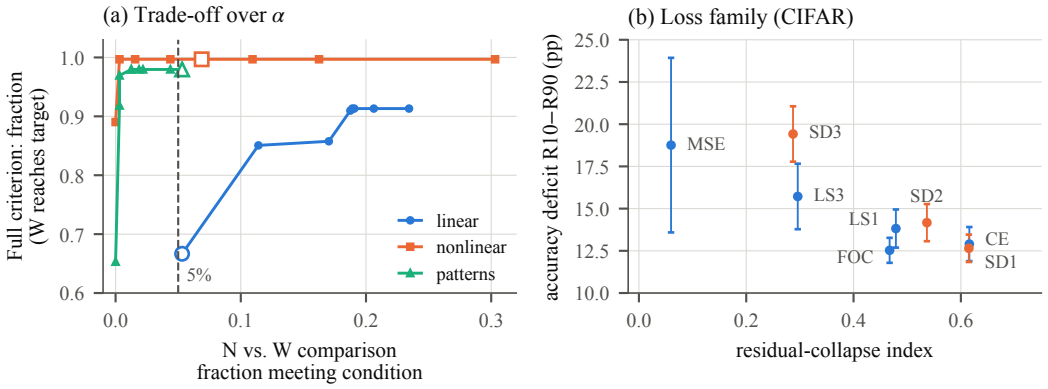


Figure 4: **(a)** Exploratory threshold comparisons on confirmation records. Both axes show fractions meeting the indicated condition; the full criterion uses only records where W reaches the target, with the subsets of Table 1. Open markers show the frozen thresholds and the dashed line marks 5%. **(b)** CIFAR accuracy deficit versus residual-collapse index across eight loss configurations (4 seeds, 95% t-intervals). Orange denotes spectral decoupling. LS1/3 use label smoothing 0.1/0.3; SD1/2/3 use strengths $10^{-3}/10^{-2}/10^{-1}$; FOC is focal loss. The positive deficit is fresh-head accuracy on random-code inputs in R10 minus R90.

confirmation frequencies of 11.4% in F-lin and 10.9% in F-nl. At $\alpha = 0.30$, F-nl has 4.4% control frequency and 0.997 full-criterion frequency; at $\alpha = 0.50$, its control frequency is 0.3%. F-e2r at $\alpha = 0.15$ has 2.2% control frequency and 0.980 full-criterion frequency. F-lin at $\alpha = 0.50$ has 11.4% and 0.851, respectively. Full-criterion frequencies use the same subsets as Table 1. These exploratory choices do not replace the frozen thresholds or validate a new operating rule (Figure 4a).

E.2 Response measurements and replications

Changing the measurement (A4, A5). Moving synthetic probes onto the training-data distribution leaves the decisions unchanged in 320/320 records. Four CIFAR response measures agree in only 5/8 seeds. Across injected-pattern checkpoints, the median Spearman correlation between the chosen response and weak-only accuracy is 0.786 for W and 0.791 for B. Reversing the roles of the strong and weak features yields no positive result under the tested criterion (0/96).

Earlier synthetic comparisons. In the first replication's linear grid at lags 0–4, 77/96 N-versus-W runs show some negative difference, but only 16/96 meet the fixed minimum-deficit condition (all at $\rho = 8$). The B-versus-W condition holds in 84/96 runs. In the tanh setting, all 8/8 seeds show a negative N-versus-W difference both somewhere during training and at the target crossing; 2/8 meet the minimum-deficit condition.

Earlier CIFAR comparisons. A separately frozen 8-seed R90-versus-W0 study (seeds 33001–33008) found original-head accuracy gaps of -19.9 points without the cue and -31.9 under cue conflict, and a fresh-head gap of -10.8 points. That study lacked an uninformative-cue comparison.

In the later fixed-threshold study, the B-versus-R10 condition at $\alpha = 0.10$ holds in 0, 0, 2, 8, and 8 of 8 seeds for cue reliabilities 0.3, 0.5, 0.7, 0.9, and 1. R30 versus R10 shows some negative difference in 8/8 seeds despite an accumulated normalized deficit of only 0.014.

Cue presence and architectural replications. The R10-minus-W0 difference in fresh-head cue-free accuracy is $+0.13$ $[-0.50, +0.76]$ points; its interval includes zero. R90-versus-R10 deficits also occur without weight decay (-13.8 points), with ResNet-18 (-15.1), AdamW (-10.0), BatchNorm (-10.0), and an MNIST–CIFAR domino construction adapted from Shah et al. (2020) (our deficit: -23.0). The full criterion is met in 8/8 seeds for both AdamW and BatchNorm.

Layer probes and weight replacement. Stitching W0's first k blocks onto R90's remaining blocks and head gives fresh-head cue-free accuracy 0.737, 0.643, 0.617, 0.712, and 0.848 for $k = 0, \ldots, 4$. The signed R90-minus-W0 probe gap is 9.4 [8.0, 10.8] points higher at block 2 than at block 4, meaning a smaller deficit at block 2. Replacing B's weak-input weights with W's raises B's weak response in 412/413 synthetic and injected-pattern records that meet the suppression criterion.

A reference prediction under cue conflict. Under a conflicting cue, R90 reaches 52.7% accuracy. A model that combines W0's calibrated class probabilities with the known cue likelihood reaches 71.4%, a difference of 18.7 [17.3, 20.1] points. This is a model-based reference, not a proved Bayes optimum. Appendix D describes its assumptions.

Additional Waterbirds comparisons. WB, WR, and WG denote training with official, independent, and gray backgrounds. Per seed, the fresh-head WB-minus-WR gray-background accuracy difference ranges from -0.8 to -0.1 points and WB-minus-WG from -1.5 to -1.1. Neither the full criterion nor the N-versus-W condition is met (0/8 and 0/8, respectively). Making background and label independent by reweighting gives the same decision as changing background pixels in 8/8 seeds; the correlation between their normalized deficits is 0.374.

E.3 Cue integration and loss functions

Adding the cue at the output. In late-fusion CIFAR, the cue enters through a separate, zero-initialized 10×10 logit map, bypassing the encoder. Its fresh-head cue-free accuracy deficit is 0.432 [0.399, 0.465] times the early-fusion deficit (8 seeds). This comparison shows that cue integration matters. It does not identify what fraction of the deficit is caused by either term of the response decomposition: Proposition 3(i) gives $\mathrm{SW}(0) = \mathrm{G}(0) = 0$, but later sensitivities can differ. It also measures fresh-head accuracy rather than the original-head response used in the decomposition. Synthetic late fusion has zero initial rate difference; transfer followed by suppression occurs in 0/192 runs. The absence of transfer in these runs is consistent with the theorem, although the noisy experiments lie outside its noiseless assumptions.

Loss functions. The positive deficit is R10-minus-R90 fresh-head accuracy on random-code inputs. The residual-collapse index measures the reduction in the accumulated output loss gradient relative to W0 (Appendix D). Across eight loss configurations, the Spearman correlation between the accuracy deficit and this index is -0.69 (permutation $p = 0.068$; Figure 4b). This association is inconclusive at the reported sample size. Table 5 gives the individual estimates. These are configuration-specific differences. They do not show that every spectral-decoupling setting harms learning or that a larger paired deficit means lower absolute B accuracy.

Contributions to the response difference. The integrated ratio ϕ_{SW}, defined in Appendix A, is -2.79 $[-3.47, -2.19]$ for linear RNNs, where the two terms have opposite signs; 0.40 [0.38, 0.42] for tanh; 0.43 [0.42, 0.45] for GRU; 0.62 [0.56, 0.78] for injected-pattern MNIST; and 0.63 [0.57, 0.72] for Fashion-MNIST. CIFAR expected-drift ratios are 0.29 [0.13, 0.54] for R90 versus W0 and 0.38 [0.20, 0.61] for R90 versus R10. These are means with 95% intervals, and are signed contributions rather than causal percentages. The full-batch CIFAR tests of early and late dominance are inconclusive: early 325 $[-272, 922]$, late $-6{,}883$ $[-23{,}420, 9{,}655]$.

Exchanging parts of the update. The crossed-field comparison exchanges output loss gradients and parameter sensitivities between cue conditions, as specified in Appendix D. The response contrast

Table 5: CIFAR loss comparisons (4 seeds). The deficit is fresh-head R10-minus-R90 accuracy on random-code inputs, in pp, with 95% t-intervals. The last column measures the reduction in accumulated residual norm relative to W0.

Loss	Accuracy deficit [95% interval]	Residual-collapse index
Cross-entropy	12.9 [11.9, 13.9]	0.62
Label smoothing, 0.1	13.8 [12.7, 15.0]	0.48
Label smoothing, 0.3	15.7 [13.8, 17.7]	0.30
Spectral decoupling, 10^{-3}	12.7 [11.8, 13.5]	0.61
Spectral decoupling, 10^{-2}	14.2 [13.1, 15.3]	0.54
Spectral decoupling, 10^{-1}	19.4 [17.8, 21.1]	0.29
Focal loss	12.5 [11.8, 13.3]	0.47
Mean squared error	18.8 [13.6, 23.9]	0.06

$C = M(\mathrm{BW}) - M(\mathrm{WB})$ is -0.62 $[-1.32, 0.07]$ on full-batch CIFAR (4 seeds), with an interval including zero. In synthetic linear models it is 0.26, 0.32, and 0.36 at 0.1, 0.25, and 0.5 of the training budget; in the injected-pattern models it is 70, 80, and 94 at steps 100, 200, and 500. In these synthetic and injected-pattern interventions, the update using cue-on sensitivities and cue-off loss gradients restores more weak response. The median relative error of the one-step first-order approximation is 0.006.

Small and large feature changes. Kernel and lazy-training analyses describe settings where parameter changes leave the effective feature map nearly fixed (Jacot et al., 2018; Chizat et al., 2019). Suppression occurs in 8/8 linear-RNN runs in both our tested lazy setting (output scale 10, width 2048) and the feature-learning setting. The conditional frozen-tangent bound in Appendix B.8 is uninformative in every tested case (0/21 informative). It therefore does not explain the observed suppression.

E.4 Training duration and prediction checks

Long-time fits and plateaus. At $\rho = 1, 2, 4$, the median fitted reliable-cue ratios are 0.503, 0.203, and 0.0586. On the separate noisy-label grid, 160/160 predicted target crossings agree with the observations, but the grid contains no predicted failures. In injected-pattern MLPs with weight decay 10^{-2}, W plateaus below the target in 64/64 records. This is an empirical observation for MLPs, not an application of the theorem for models linear in their parameters.

Longer CIFAR training and cue changes. Between the two tested training budgets, the average R90 fresh-head cue-free accuracy recovery is 1.8 [0.5, 3.0] points per doubling, and the difference between R90 and R100 slopes is 2.7 [0.8, 4.5]. These are two-budget summaries. After 471 updates following a cue change at 0.1, 0.25, or 0.5 of training, removal gives 6.0 [4.0, 8.0], 7.6 [6.6, 8.7], and 5.3 [4.5, 6.2] points of recovery. Replacing the cue with an uninformative one gives 7.7 [6.9, 8.5], 7.6 [6.5, 8.8], and 4.8 [4.2, 5.4]. Adding the cue to a cue-free run gives -6.8 $[-7.6, -6.0]$, -3.1 $[-3.8, -2.4]$, and -0.8 $[-1.5, -0.1]$. All compare with the corresponding unchanged branch (8 seeds, 95% t-intervals).

Initial learning rates (A2, C1–C4). Across 46,080 noiseless linear-RNN records, the initial-rate formula matches autograd to 10^{-9} and agrees with the early difference's sign in 46,000/46,078 evaluated comparisons (0.998). Early N-versus-W signs agree with $-\mathrm{sign}(Ks)$ in 94/96 noisy linear-RNN records, 365/384 replication records, and the tested tanh and GRU records. The latter are empirical extensions of the noiseless identity. For predicting the $H = 5$ outcome, the initial rate has AUROC 0.617 overall; by width 32, 96, and 256, it is 0.649, 0.616, and 0.579, and by gain 0.3, 0.5, 0.8, and 1.0, it is 0.578, 0.599, 0.628, and 0.656. The difference changes sign before H in 45.4% of records. With noise and nonlinearity, the autograd rate gives AUROC 0.707 versus 0.586 for the noiseless formula, a gain of 0.121 [0.084, 0.158]. The formula describes the initial rate, not a general forecast of later suppression.

In the lag-profile comparison, the lag with the largest mean normalized deficit at $H = 5$ is 6, 12, 16, and 0 for gains 0.3, 0.5, 0.8, and 1.0. The negative initial rate peaks at lag 3 for every gain;

correlations across lags are 0.65, 0.30, -0.12, and 0.87. Across cue reliabilities, the per-initialization slope of normalized deficit on $2r - 1$ is 0.659 [0.635, 0.683].

Forecasting a later deficit. The prespecified forecaster based on accumulated rates (Appendix B.8) is compared with linear extrapolation on 675 records. The paired prediction-error difference is -0.015 $[-0.034, 0.005]$; the interval does not establish an improvement. Among records meeting the criterion, a period of transfer precedes suppression in 0.381 of synthetic and 0.979 of injected-pattern records. A non-rigorous plug-in version of Proposition 4 predicts outcomes less accurately than the sign of the current difference (-0.148 $[-0.179, -0.115]$).

E.5 Attempts to restore weak-feature learning

On injected-pattern data, spectral decoupling restores less weak response in runs meeting the suppression criterion than in the low-amplitude comparisons (differences -0.046 and -0.070). Upweighting misclassified examples has no effect because B classifies every training example correctly. Refitting the last layer on inputs with an uninformative cue restores more response in runs meeting the criterion, measured in raw logit units. These different response scales should not be interpreted as comparable accuracy gains.

Table 6 compares the local response correction of Appendix B.9 with an unconstrained correction. Preserving the strong-response derivative at one point does not preserve the strong response over a complete training trajectory. The finite-time results therefore do not support this correction as a remedy.

Table 6: CDC compared with plain B training and with unconstrained rescue $v+(\delta/\|p\|^2)p$ (synthetic, 16 seeds, mean [95% CI]). The correction ratio is $\|\gamma q\|/\|v\|$ (median; max). The last column is the fraction of steps with $\cos(p, a) < 0$.

Arch.	Corr. ratio	ΔM_s CDC	ΔM_s rescue	ΔM_w CDC	$\cos < 0$
linear	8.86; 40.6	-0.377 [-0.406, -0.348]	-0.118 [-0.148, -0.088]	3.24 [3.01, 3.48]	0.085
tanh	3.12; 540	-0.762 [-0.905, -0.619]	-0.323 [-0.374, -0.271]	5.67 [5.36, 5.97]	0.019
GRU	8.53; 19.6	-0.296 [-0.316, -0.276]	$+0.315$ [0.276, 0.354]	3.63 [3.51, 3.75]	0.189

E.6 DETAILED NUMERICAL BOUNDS

The verified nonnegative-gap fraction is 13–28% across the nine network configurations. For D1's 16 initializations per configuration, the largest verified relative deficits range from 0.31 to 1.20. Suppression counts are unchanged for the tested shared initialization boxes of radius up to 10^{-4}. The implementation checks reproduce the single-ball construction, contain a high-precision reference trajectory on each configuration, and reject a false claim. Those checks test consistency; the rigorous conclusions rely on the interval enclosures themselves.

Table 7: Rigorous B-versus-W bounds (Theorem 3). D1: 16 initializations per configuration. *Suppressed* denotes the validated inequality $\Delta_{\mathrm{B|W}} \leq -0.1M^{\mathrm{W}}$ on $[H/2, H]$ plus the W target crossing; H is chosen separately for each initialization as in Appendix D. Suppressed counts are unchanged under shared initial boxes of radius 10^{-8}, 10^{-6} and 10^{-4}. Columns S/N/U denote suppressed, nonnegative and undetermined. Max rel.: largest verified relative deficit; CPU s: median runtime in seconds. D2: fractions over 1,000 initializations from each specified initialization law, with Clopper–Pearson (CP) 95% intervals for the suppressed fraction. *Nonnegative* excludes a negative gap on that interval. The inequality expresses relative suppression where $M^{\mathrm{W}} > 0$. These are not full-criterion results.

			D1: 16 initializations			D2: 1,000 initializations		
ID	Model	P	S/N/U	Max rel.	CPU s	S (CP 95%)	N	U
L0	tanh CNN	5	6/4/6	0.62	1.3	.408 [.377,.439]	.279	.313
L1	linear RNN	18	5/5/6	0.70	0.8	.419 [.388,.450]	.239	.342
L2	linear RNN	10	6/4/6	1.20	0.6	.418 [.387,.449]	.269	.313
L3	linear RNN	28	7/6/3	0.77	3.7	.408 [.377,.439]	.239	.353
L4	linear RNN	54	5/3/8	0.60	13.6	.406 [.375,.437]	.214	.380
L5	tanh RNN	32	7/3/6	1.17	16.0	.506 [.475,.537]	.139	.355
L6	tanh MLP	49	7/1/8	0.63	3.4	.493 [.462,.524]	.138	.369
L7	tanh MLP	97	9/2/5	1.02	36.7	.484 [.453,.515]	.129	.387
L8	tanh MLP	145	5/3/8	0.31	166.9	.417 [.386,.448]	.129	.454

F PRE-REGISTERED ENDPOINT LEDGER

Table 8: Every pre-registered primary endpoint of the second study, with its frozen rule and historical status. Terminology is clarified in the text without changing any numerical decision. *Geometry* for B2-P1 is the original threshold-based label for the fusion ratio, not an identified causal mechanism.

Endpoint	Quantity (rule)	Estimate	Status
A1-P1	N|W control frequency at $\alpha^\star$ (≤ 0.05, upper ≤ 0.10)	0.053, 0.069, 0.053; 0/8	not met (3 fam.)
A1-P2	full criterion rate among B records with W reaching the target (A1-P2 subset) (≥ 0.75)	0.667, 0.997, 0.980, 1.000	met (3 of 4)
A2-P1 / A2-P2	N|W sign law (≥ 0.80) / score AUROC (lower > 0.5)	94/96 / 0.306	met / not met
A3-P1 / A3-P2	asymptotic ratio fits / predicted target crossing	96/96 / 160/160	met / met
A3-P3 / A3-P4	E2R a_1 target crossing by $H = 5000$ (≥ 0.90) / plateau (report)	0/64 / 64/64	not met / reported
A4-P1 / A4-P2	probe invariance (≥ 0.90) / measure invariance ($\geq 7/8$)	320/320 / 5/8	met / not met
A5-P1	reversed role (≤ 0.05)	0/96	met
B1-P1	sample-weight share, full-batch CIFAR (report)	-0.083 [-0.445, 0.100]	reported
B1-P2	early geometry / late sample-weight dominance	SYN 0.183 / -0.342; CIFAR n.s.	met / not met
B1-P3	remainder check, full-batch CIFAR	0/8	not met
B2-P1 / B2-P2	fusion ratio (decision rule) / late-fusion transfer (≤ 0.05)	0.432 / 0/192	geometry / met
B3-P2	loss-family rank correlation ($\|\rho\|$ small)	-0.69 ($p = 0.068$)	not supported
B4-P1	crossed-field contrast ($C < 0$, upper < 0)	-0.62 [-1.32, 0.07]	not met
B5	closure checks	pass	met
B6-P1	drift-balance forecaster vs. baseline (upper < 0)	-0.015 [-0.034, 0.005]	not met
B7-P1 / B7-P2	SGD band coverage (≥ 0.95) / variance ratio (upper < 1)	4/16 / 0.984	not met / not met
C1-P1 / C1-P2	early sign (≥ 0.99) / AUROC (lower > 0.5)	0.998 / 0.617	met / met
C3-P1	argmax lag matches ($\geq 3/4$ gains)	0/4	not met
D1-P1	bounds for configurations L2–L4	6, 7, 5 inits	met
D3-P1	plug-in forecast vs. baseline (lower > 0)	-0.148	not met
D4-P1 / D5-P1	transplant gain (≥ 0.9) / layer gap (lower > 0)	0.998 / 0.094	met / met
E1-P2	lazy-corner suppression (≥ 0.75)	8/8	met
F1	CDC local guarantees	60/60	met
H1-P1 / H1-P2	Waterbirds suppression criterion (≥ 0.75) / WB$-$WR (upper < 0)	0/8 / -0.005	not met / met
H2-P1	reweighting agreement (≥ 0.75)	8/8	met
H3-P1 / H3-P2	R90 slope / R90$-$R100 slope (lower > 0)	0.018 / 0.027	met / met
H4a / H4b	AdamW / BatchNorm (upper < 0)	-0.100 / -0.100	met / met
H5-P1	cue removal at $0.25H$ (lower > 0)	0.076	met
H6-P1	remedy recovery criterion met $-$ not met (lower > 0)	SD -0.046, -0.070; DFR 23.1	DFR met

Theory checks are in Table 3. The first replication met its CIFAR and E2R endpoints; 16.7% N|W control frequency left its fixed-threshold rule undecided and motivated calibration. Its residual-swap test diverged under an unbounded swapped loss and is not used.