
Beyond Accuracy and Alignment: A Diagnostic Evaluation Protocol for Feedback Alignment

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Modern feedback-alignment evaluation on deep residual networks is still summar-
2 ized by a deceptively simple pair: headline accuracy and headline cosine align-
3 ment Γ to the backpropagation gradient. We show that this pair can silently fail in
4 two distinct ways on standard CIFAR-10 pre-LayerNorm ResMLP and ViT-Mini
5 settings: first, *measurement degeneracy*, where residual-stream growth drives
6 hidden-layer BP gradients to the numerical floor and makes Γ uninterpretable;
7 and second, *low intrinsic credit-direction quality*, where random-feedback credit
8 remains essentially unaligned with BP on the deep blocks even when the reference
9 gradient is still meaningful. The headline result is that the field-standard reporting
10 pair walks back none of the methods we audit, whereas a four-diagnostic proto-
11 col walks back the three degenerate methods and passes the two trustworthy con-
12 trols. Intervention with a per-block scale-control penalty further reveals method-
13 dependent severity within the audited fixed-feedback family: State Bridge then
14 exceeds the architecture-matched frozen-blocks baseline by about 10 percentage
15 points, while Credit Bridge attains much higher deep BP cosine than DFA at the
16 same final accuracy, a dissociation that motivates reporting layerwise credit quality
17 jointly with a depth-utilization baseline. Our contribution is an evaluation method-
18 ology paper for the NeurIPS 2026 Evaluations & Datasets track: we provide the
19 protocol, the calibration logic for its thresholds, a reference implementation, a five-
20 method audit, and validation through temporal replay, cross-architecture checks,
21 intervention-based disambiguation, and a documented catalog of pipeline pitfalls,
22 in the spirit of critical evaluation analyses such as Jordan et al. [3], O’Bray et al.
23 [2], Paleka et al. [1].

24 1 Introduction

25 Backpropagation (BP) is the de facto training method for deep neural networks, but its requirement
26 that each feedback connection carry a weight identical to the corresponding forward connection –
27 the weight-transport problem – has long been considered biologically implausible [4, 8]. *Feedback*
28 *alignment* (FA) [4] side-steps weight transport by delivering per-layer credit through fixed random
29 feedback matrices, and its direct variant (DFA) [5] projects the output error to every hidden layer
30 through an independent random matrix; parallel lines include target propagation [15] and equilib-
31 rium propagation [9]. These rules are studied both as biologically-plausible alternatives to BP and
32 as scalable, asynchronous training schemes, with recent work scaling DFA to transformer-scale ar-
33 chitectures on language, recommendation, and view-synthesis tasks [7, 6]. Evaluation in this line of
34 work has converged on a two-number summary: final task accuracy, and an aggregate cosine align-
35 ment Γ between the method’s per-layer credit and the BP gradient on the trained network [4–8].

36 On the audited 4-block $d=256$ ResMLP, however, Table 1 already shows that this accuracy-plus- Γ
 37 pair is not a validity check: DFA reaches only 0.306 ± 0.006 test accuracy, below the architecture-
 38 matched frozen-blocks baseline of 0.349 ± 0.002 , while still looking superficially comparable to
 39 other non-BP methods. Figure 1 further shows that the apparent cosine evidence is concentrated
 40 at the shallowest block, with DFA at seed 42 reaching about $+0.42$ at layer 0 but approximately
 41 -0.03 to 0 on layers 1–4, so the aggregate obscures where credit direction is and is not present.
 42 At the same time, the deepest BP reference norm is only about 4×10^{-10} for DFA (three-seed
 43 mean) and a few $\times 10^{-9}$ for State Bridge and Credit Bridge, all below the 10^{-8} clamp used by
 44 `F.cosine_similarity`, whereas BP remains around 4×10^{-4} , so the reported deep cosine is
 45 partly computed against a numerical-floor reference rather than an informative gradient direction
 46 (Figure 1; Table 1). Those numbers can be useful, but only if the measurement regime itself is valid.

47 Our audit shows that modern residual vision models can make these two quantities look informa-
 48 tive while failing to answer the question they are taken to answer. Figure 1 shows the first failure
 49 mode, which we call *Mode 1: measurement degeneracy*, where residual-stream growth drives the
 50 deepest hidden state to about $\|h_L\| \sim 10^8$ under DFA/SB/CB while the corresponding BP reference
 51 collapses to $\|g_L\| \sim 4 \times 10^{-10}$ for DFA (three-seed mean), so the deep-layer cosine is measured
 52 against a clamp-dominated floor rather than a meaningful target direction. The same figure also
 53 shows the second failure mode, *Mode 2: low intrinsic credit-direction quality*, because even after
 54 comparing against the stronger frozen-blocks baseline (0.349 ± 0.002) and looking layer-by-layer,
 55 DFA’s deep blocks remain essentially null while only layer 0 is visibly positive. Intervention sharp-
 56 ens both modes. Adding a per-block residual penalty $\lambda \|f_l(h_l)\|^2$ to DFA at $\lambda=10^{-2}$ contains $\|h_L\|$
 57 to about 4×10^4 and lifts the deep BP reference to about 10^{-6} , but DFA’s rescued deep cosine is
 58 only about $+0.16$; State Bridge under the same intervention reaches a three-seed deep cosine of
 59 $+0.32$ and, unlike DFA, exceeds the frozen-blocks baseline by $+10$ points in final accuracy; Credit
 60 Bridge reaches a deep cosine near $+0.68$ yet matches only the DFA accuracy, so Mode 2 has method-
 61 dependent severity and deep cosine is not a sufficient predictor of final accuracy across methods. At
 62 the same time, at $\lambda=10^{-4}$ Mode 1 is alleviated while the DFA deep cosine still stays near zero, and
 63 at vanilla DFA epoch 1 the reference is already meaningful at about 6×10^{-7} but the deep cosine is
 64 still -0.008 ± 0.013 across three seeds. The failure is therefore neither unitary nor uniform: Mode 1
 65 and Mode 2 are observationally separable, and within the audited fixed-feedback family, the severity
 66 of each mode varies by method.

67 Accordingly, this paper does not introduce a new FA variant or a new benchmark. Of the five
 68 methods we audit, BP, EP, and DFA are established baselines from the published literature; the
 69 remaining two, which we call *State Bridge* and *Credit Bridge*, are diagnostic probes we construct
 70 in this paper to directly learn the two targets that different strands of the BP-free literature argue
 71 should produce good per-layer credit (formal definitions and citations in Section 2). Instead, Table 1
 72 and Figure 1 use a standard five-method CIFAR-10 audit to show that status-quo reporting would
 73 treat BP, EP, DFA, State Bridge, and Credit Bridge as the same kind of evidence-bearing object
 74 even though only BP and EP remain trustworthy under matched diagnostic checks. This makes the
 75 contribution methodological in the sense of Jordan et al. [3], O’Bray et al. [2], and Paleka et al. [1]:
 76 the central question is not whether one more FA variant can post a headline number, but whether the
 77 reporting pipeline distinguishes meaningful credit-direction evidence from numerical-floor artifacts
 78 and from shallow-only learning. The protocol therefore starts from per-layer diagnostics and a
 79 frozen-blocks baseline before reading any aggregate cosine or final accuracy as evidence about deep
 80 credit assignment. We first show the walk-back on a standard audit, then isolate the two failure
 81 modes, and finally state the reporting protocol that future FA papers should satisfy.

82 2 Audit: Standard Reporting Walks Back Nothing

83 Table 1 fixes the canonical audit to a 4-block pre-LayerNorm ResMLP with width $d=256$ on CIFAR-
 84 10, trained for 100 epochs with AdamW (learning rate 10^{-3} , weight decay 0.01), a cosine schedule,
 85 batch size 128, and three seeds (42, 123, 456); all five methods are read against the identical ar-
 86 chitecture, optimizer, schedule, and training budget without method-specific tuning, and Figure 1
 87 summarizes the corresponding per-block growth, deepest-layer BP reference norm, cross-batch sta-
 88 bility, and frozen-baseline comparison.

89 Two rows in Table 1, *State Bridge* (SB) and *Credit Bridge* (CB), are diagnostic probes we
 90 construct in this paper, not prior FA variants. Each directly learns a target that a different

Table 1: Main audit table for the 4-block $d=256$ pre-LayerNorm ResMLP on CIFAR-10. The row and column structure is fixed here; fill from the three-seed audit output.

| Method | Test acc. | Headline Γ | Status-quo verdict | Protocol verdict |
|---------------|-------------------|-------------------|--------------------|------------------|
| BP | 0.615 ± 0.003 | ≈ 1.0 | trustworthy | trustworthy |
| EP | 0.316 ± 0.030 | 0.008 | trustworthy | trustworthy |
| DFA | 0.306 ± 0.006 | 0.10 | trustworthy | walked back |
| State Bridge | 0.205 ± 0.032 | 0.005 | trustworthy | walked back |
| Credit Bridge | 0.289 ± 0.026 | 0.07 | trustworthy | walked back |

91 strand of the BP-free literature argues should produce good per-layer credit, and each uses the
 92 same block local loss $-\langle f_l(h_l), a_l \rangle$ as DFA but with a different a_l . SB instantiates the target-
 93 propagation view that accurate prediction of a downstream hidden state yields a usable credit
 94 signal [14, 15]: an auxiliary $G_\psi(h_l, t_l, s)$ is fit by MSE to predict h_L from $(h_l, t_l=l/L, s=e_T)$,
 95 and $a_l^{\text{SB}} = \nabla_{h_l} \text{CE}(W_{\text{out}} \text{LN}(G_\psi(h_l, t_l, s)), y)$. CB instantiates the synthetic-gradient view that a
 96 learned value network, if its input-gradient approximates the BP gradient, can stand in for it [16]:
 97 $V_\phi(h_l, t_l, s)$ is fit via a bridge residual against an EMA target, and $a_l^{\text{CB}} = \nabla_{h_l} V_\phi(h_l, t_l, s)$. Both
 98 auxiliaries are trained on detached hidden states. We use SB and CB as controls that populate differ-
 99 ent points in the (angular agreement with BP, functional usefulness) plane; that is what makes the
 100 cross-method cosine-versus-accuracy dissociation in Section 4 visible.

101 By the field’s usual criteria, the non-BP methods appear to train to nontrivial accuracy and report
 102 nonzero alignment. In Table 1, DFA reaches 0.306 ± 0.006 test accuracy with headline $\Gamma=0.10$,
 103 State Bridge reaches 0.205 ± 0.032 with $\Gamma=0.005$, and Credit Bridge reaches 0.289 ± 0.026 with
 104 $\Gamma=0.07$; none of these rows looks like an obvious invalidation if one is reading the usual pair of final
 105 accuracy and aggregate alignment in the style of prior FA reporting [4–7]. Even the absolute scale
 106 does not itself force a walk-back, because all three methods are plainly above chance and all three
 107 report positive headline alignment rather than a visibly broken or undefined quantity. That reading
 108 is exactly what the rest of the paper overturns.

109 Low accuracy by itself is not the pathology. Equilibrium Propagation (EP), a contrastive energy-
 110 based alternative to BP that updates weights from the difference between a free-phase and a nudged-
 111 phase hidden trajectory, is the key internal comparison in Table 1 and Figure 1: it achieves only
 112 0.316 ± 0.030 accuracy and a very small headline $\Gamma=0.008$, yet its three-seed mean max-per-block
 113 growth is only $6.6\times$ (highest single-seed value $11.0\times$), its deepest BP reference norm remains
 114 around 1.3×10^{-4} rather than collapsing to the numerical floor, and its cross-batch direction-stability
 115 score is 0.02 rather than the much higher drift-dominated values seen for DFA-family methods. At
 116 the same time, EP is not a positive result for depth usage in the stronger sense, because its trainable-
 117 model accuracy is still 3.3 percentage points below the frozen-blocks baseline of 0.349 ± 0.002 . The
 118 distinction matters because it separates underperformance from invalid evaluation.

119 When we compare each method to a frozen-blocks baseline matched to the same architecture, the
 120 headline interpretation changes immediately. The frozen-blocks model, which trains only the em-
 121 bedding, LayerNorm, and head while holding the residual blocks fixed, reaches 0.349 ± 0.002 across
 122 the same three seeds; against that baseline, BP is higher by 26.6 points, but DFA is lower by 4.3
 123 points, State Bridge by 14.4 points, Credit Bridge by 6.0 points, and even EP by 3.3 points. Fig-
 124 ure 1 shows that this accuracy comparison lines up with the diagnostic split: DFA, State Bridge,
 125 and Credit Bridge also combine extreme per-block growth (three-seed mean max ratios $\sim 1.9 \times 10^3$,
 126 $\sim 1.6 \times 10^4$, and $\sim 1.2 \times 10^3$ respectively), deepest-layer BP norms around 10^{-9} , and high cross-
 127 batch instability (0.16, 0.53, and 0.37), so their deep blocks are at best passengers and in practice
 128 often harmful. This establishes the audit question the rest of the paper must answer: why do the
 129 standard signals fail so badly?

130 3 Failure Mode 1: Measurement Degeneracy

131 Mode 1 has two parts. The activation-growth part (a) is a scale pathology of fixed-feedback local-
 132 credit objectives without an effective scale-control term: for block l , DFA, State Bridge, and Credit

5-method audit on 4-block $d=256$ ResMLP CIFAR-10 (3-seed mean \pm std)

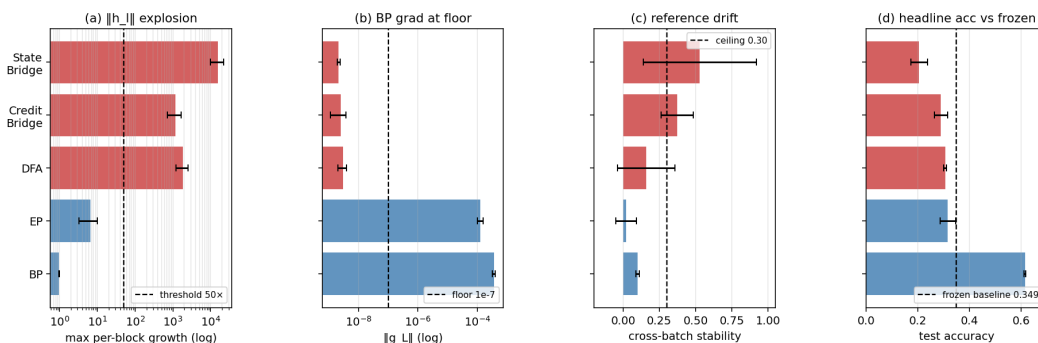


Figure 1: Five-method audit on the 4-block $d=256$ pre-LayerNorm ResMLP: the field-standard pair looks superficially consistent across methods, but the diagnostic view separates trustworthy controls from walked-back methods.

133 Bridge each update f_l by maximizing a local objective of the form $\langle f_l(h_l), a_l \rangle$, where the per-layer
 134 credit vector a_l is the method-specific projection of the output error (for DFA, $a_l = B_l^\top e_T$ with
 135 a fixed random B_l ; for State Bridge, a_l is the gradient of a cross-entropy loss measured through
 136 a learned state predictor $G_\psi(h_l, t_l, s)$ that estimates h_L ; for Credit Bridge, a_l is the gradient of a
 137 learned value network $V(h_l, t_l, s)$). None of these three local losses contains a penalty on $\|f_l(h_l)\|$,
 138 so any direction in which a larger block output improves inner-product alignment with the method’s
 139 fixed or learned credit target is rewarded; in a pre-LN residual stack, larger block outputs directly
 140 increase residual-stream scale, and terminal LayerNorm at the output removes task-loss sensitivity
 141 to that scale, so the architecture supplies no global restraint on the local growth incentive. The
 142 gradient-floor part (b) follows from the LayerNorm Jacobian. For $y = \text{LN}(h) = (h - \mu(h))/\sigma(h)$
 143 with $\sigma(h) = (\frac{1}{d} \sum_i (h_i - \mu(h))^2)^{1/2}$ proportional to $\|h\|/\sqrt{d}$, the spectral norm of $\partial y/\partial h$ is
 144 $\Theta(1/\sigma(h))$, so back-propagating through terminal LayerNorm scales the deepest hidden BP gra-
 145 dient as $\|g_L\| = \Theta(1/\|h_L\|)$, and the same residual-stream inflation that drives diagnostic (a) drives
 146 a proportional collapse of the diagnostic (b) reference. Empirically, on the audited 4-block pre-
 147 LayerNorm ResMLP ($d=256$, CIFAR-10, 100 epochs, 3 seeds), DFA training drives the three-seed
 148 mean $\|h_L\|$ from about 9 at initialization to about 5×10^8 by epoch 100 and $\|g_L\|$ from about
 149 9.8×10^{-4} to about 4×10^{-10} , while the reported deep cosine remains defined only because
 150 `F.cosine_similarity` clamps the denominator at $\epsilon=10^{-8}$ (Table 1; Figure 1). At that endpoint
 151 the reference norm is about $20\times$ below the clamp, so the quantity being reported is effectively
 152 $(a \cdot b)/(\|a\| \max(\|b\|, 10^{-8}))$ rather than a comparison to a meaningful BP direction.

153 We tested this mechanism story against four natural alternative attributions, all of which it survives.
 154 *Not residual-skip-driven*: with terminal LN kept and the additive skip removed ($h_{l+1}=F_l(h_l)$),
 155 DFA still converges across three seeds to mean $\|h_L\| \approx 8.2 \times 10^7$ and mean $\|g_L\| \approx 1.9 \times 10^{-10}$ at 100
 156 epochs, both at the diagnostic floor (Appendix H). *Not task-signal-driven*: under i.i.d. random class
 157 targets per minibatch, DFA still reaches $\|h_L\| \approx 1.67 \times 10^8$ and $\|g_L\| \approx 8 \times 10^{-12}$ while accuracy stays
 158 at chance (Appendix I). *Not DFA-specific*: the same random-target ablation drives $\|h_L\|$ to 6.2×10^3
 159 for SB and 2.0×10^4 for CB in three epochs, so all three audited fixed-feedback methods exhibit
 160 data-agnostic activation growth. *Not shared by EP*: under the same protocol, EP keeps $\|h_L\| \approx 586$
 161 at five epochs, $25\times$ smaller than DFA’s three-epoch value, confirming that the random-target assay
 162 separates the explosion-prone fixed-feedback class from EP’s energy-based objective.

163 The matched same-backbone causal control for diagnostic (b) is removing terminal LayerNorm. On
 164 the same ResMLP-d256 with the residual skip intact, 100 epochs of DFA, three seeds, the resid-
 165 ual stream still inflates to $\|h_L\| \approx 1.21 \times 10^7$, but the deepest hidden-layer BP gradient remains
 166 at $\|g_L\| \approx 7.2 \times 10^{-4}$ (four orders of magnitude above the diagnostic (b) floor), and the final
 167 test accuracy is 0.327 ± 0.012 , statistically indistinguishable from vanilla DFA’s 0.306 ± 0.006
 168 on the same backbone with terminal LayerNorm intact. Removing terminal LayerNorm therefore
 169 preserves Mode 1 (a) but cleanly eliminates Mode 1 (b) on the same architecture, while leaving fin-
 170 al task accuracy essentially unchanged. Combined with the broader cross-architecture pattern (the

171 no-terminal-LN ResMLP-d256 ablation and the BatchNorm CNN, which lack terminal LayerNorm,
 172 never trigger diagnostic (b); ViT-Mini with a terminal LN does, by epochs 2–3 (Figure 2)), terminal
 173 LayerNorm is necessary for Mode 1 (b) in the audited residual ResMLP and ViT-Mini setting. The
 174 collapse is also not a late-epoch curiosity: $\|g_L\|$ drops from 9.8×10^{-4} at epoch 0 to 5.8×10^{-8}
 175 by epoch 4 in the three-seed temporal replay (per seed: 6.8, 6.4, 4.1×10^{-8}), so the protocol fires
 176 within the first 11 epochs of a 100-epoch run and is actionable as an early-stop criterion rather than
 177 a post hoc explanation. Once measurement degeneracy is identified, the next question is whether
 178 poor deep credit remains even before collapse.

179 4 Failure Mode 2: Low Intrinsic Credit-Direction Quality

180 The second failure mode appears even in the meaningful-measurement regime. At the earliest vanilla
 181 DFA checkpoints on ResMLP, the hidden backpropagated gradient at the first deep block remains
 182 above the numerical floor: at epoch 1, $\|g_2\|$ is 6.8×10^{-7} , 6.6×10^{-7} , and 3.8×10^{-7} across the three
 183 seeds, all above the 10^{-7} threshold used to distinguish measurable from collapsed gradients. Yet the
 184 corresponding deep-layer cosine values are already essentially null: across layers 1–4, all seed-level
 185 measurements at epoch 1 lie in $[-0.04, +0.02]$, with a three-seed mean of -0.008 ± 0.013 , and
 186 by epoch 2 the deep mean is still only -0.018 ± 0.018 (Table 2). This is the observational pattern
 187 predicted by low credit-direction quality rather than mere disappearance of signal: the gradient is
 188 still present enough to measure, but the directions delivered to the deep network carry little agree-
 189 ment with backpropagation, consistent with prior concerns that alternative feedback rules can fail
 190 by supplying poor credit assignments even before full collapse [8, 10, 12, 11]. This rules out the
 191 simplest objection that the deep-layer null result is merely a byproduct of collapse.

192 A second metric with different numerical failure modes tells the same story. Cosine measures direc-
 193 tional agreement with the BP gradient, whereas the per-layer perturbation correlation ρ_l measures
 194 whether the proposed credit predicts the actual loss response: for $M=32$ unit-norm random di-
 195 rections v_m and step $\varepsilon=10^{-3}$, $\rho_l = \text{Pearson}_m(\langle a_l, \varepsilon v_m \rangle, \ell(h_l + \varepsilon v_m) - \ell(h_l))$, evaluated per
 196 sample on a fixed eval batch and then averaged. Cosine and ρ have different failure modes, espe-
 197 cially with respect to normalization and small-denominator effects. In our controls, ρ behaves as
 198 expected, with a Taylor-ceiling positive control near $+0.997$ and a random-vector negative control
 199 near $+0.006$ (Figure 3, Table 2). On vanilla DFA, deep ρ is likewise null: for the early checkpoints
 200 where the gradients remain measurable, the deep average is -0.003 ± 0.005 across seeds and epochs,
 201 and in a floor-level checkpoint it is $+0.002$, again indistinguishable from noise. The agreement be-
 202 tween cosine and ρ therefore rules out the interpretation that the null deep result is an artifact of
 203 cosine’s ε -clamp or vector normalization. The deep blocks are not just hard to measure; they are
 204 receiving weakly useful directions.

205 Per-layer reporting is therefore not cosmetic. In ResMLP under vanilla DFA, the headline aggregate
 206 alignment $\Gamma \approx 0.07$ – 0.10 can look mildly positive only because layer 0 remains strongly aligned
 207 while the deep network is not: at the same epoch-1 checkpoints where layers 1–4 are essentially zero,
 208 layer 0 has cosine $+0.42$, $+0.44$, and $+0.42$ across seeds (Table 2; per-seed values in Appendix K).
 209 The resulting average can therefore be driven by the embedding layer even when the interior blocks
 210 are effectively unaligned, so aggregate reporting obscures the very distinction needed to separate
 211 “measurement collapse” from “poor credit direction.” This layer-0 dominance is specific to the
 212 ResMLP DFA setting; on ViT-Mini DFA, all layers are near zero, which strengthens the broader
 213 methodological point that alignment should be reported per layer rather than only in aggregate. With
 214 the two modes separated observationally, the remaining question is whether intervention can move
 215 them independently.

216 Mode 2 has method-dependent severity within the audited fixed-feedback family once Mode 1 is
 217 alleviated. Applying the same per-block scale-control penalty $\lambda=10^{-2}$ that rescued DFA to State
 218 Bridge and to Credit Bridge on the same 4-block $d=256$ ResMLP backbone over 30 epochs and three
 219 seeds gives converged test accuracies of 0.453 ± 0.003 (SB) and 0.360 ± 0.003 (CB), with deep mean
 220 cosines of $+0.322 \pm 0.007$ (SB) and $+0.679 \pm 0.008$ (CB) and deep mean ρ of $+0.402 \pm 0.015$
 221 (SB) and $+0.464 \pm 0.025$ (CB), while DFA under the same intervention reaches 0.360 ± 0.001
 222 with deep cosine $+0.151 \pm 0.025$ and deep ρ $+0.080 \pm 0.011$ (Table 2; Appendix J). The State
 223 Bridge penalty rescue is roughly 24 percentage points above the vanilla State Bridge baseline of
 224 0.213 on the same architecture and, more importantly for the paper’s central walk-back, exceeds
 225 the architecture-matched frozen-blocks shallow baseline of 0.349 by $+10.4$ percentage points. State

Table 2: Two-mode validation table built around the intervention and disambiguation results.

| Condition | Deep-layer alignment signal | Measurement regime | Interpretation |
|------------------------------------|--|-------------------------------------|------------------------------------|
| Vanilla DFA, early epoch | $\overline{\text{cos}}_{deep} = -0.008 \pm 0.013, \overline{\rho}_{deep} = -0.003 \pm 0.005$ | meaningful ($\ g\ \sim 10^{-6}$) | mode 2 present without mode 1 |
| Vanilla DFA, converged | $\overline{\text{cos}}_{deep} = -0.022, \overline{\rho}_{deep} = +0.002$ | degenerate ($\ g\ \sim 10^{-9}$) | mode 1 obscures mode 2 |
| Penalized DFA, $\lambda = 10^{-2}$ | $\overline{\text{cos}}_{deep} = +0.151 \pm 0.025, \overline{\rho}_{deep} = +0.080 \pm 0.011$ | meaningful ($\ g\ \sim 10^{-6}$) | partial alleviation of both modes |
| Fresh- B null control | $\overline{\text{cos}}_{deep} = +0.002 \pm 0.022$ ($n=20$ draws) | meaningful | training-specific adaptation check |

226 Bridge with the penalty intervention is therefore the first audited non-BP method whose trained deep
 227 blocks substantively improve over an architecture-matched random-block baseline; the headline ac-
 228 curacy gap is comparable to BP+penalty’s +18.3 pp over the same shallow baseline. Neither the
 229 activation scale nor the deep BP gradient magnitude is silenced under the penalty: $\|h_L\|$ stays at
 230 302 ± 8 for SB and 5680 ± 178 for CB, with $\|g_L\|$ at $\sim 1.8 \times 10^{-4}$ and $\sim 1.9 \times 10^{-5}$ respectively,
 231 both well within the meaningful-measurement regime, so the recovered deep cosines are computed
 232 against an informative reference and not against a numerical floor. Within this rescued regime, the
 233 three methods reveal a clean cosine-versus-accuracy dissociation. Credit Bridge achieves roughly
 234 $4\times$ the deep cosine of DFA and $2\times$ that of State Bridge, yet its final accuracy matches DFA’s and
 235 is 9 percentage points below State Bridge’s. We therefore frame the Mode 2 reading as a three-part
 236 proposition. *Observation:* under the same intervention and matched training budget, CB and DFA
 237 reach the same accuracy despite a $4\times$ deep-cosine gap, while SB is the best accuracy with interme-
 238 diate cosine. *Inference:* layerwise cosine to the BP gradient is necessary to rule out grossly wrong
 239 credit signals (it distinguishes the rescued regime from the clamp-dominated vanilla regime), but
 240 it is not sufficient to certify that the supplied signal is useful credit for depth. *Mechanism hypoth-*
 241 *esis:* usefulness depends on whether the local update induces useful forward-state change across
 242 blocks, not merely whether its direction is close to the BP gradient in angle. Under this reading, CB
 243 supplies a gradient-direction surrogate that aligns with BP in angle but does not translate to a coordi-
 244 nated forward-state improvement, while State Bridge supplies a state-level downstream teaching
 245 signal that preserves aspects of useful credit which layerwise cosine does not measure. We state this
 246 as a mechanism hypothesis rather than a theorem because we have measured the angle-to-accuracy
 247 gap but not the full functional-credit content; the reporting rule that follows is robust to either inter-
 248 pretation. This cross-method dissociation strengthens the methodological point that alignment must
 249 be reported jointly with measurement validity and a depth-utilization baseline rather than as a single
 250 headline number.

251 5 Intervention and Cross-Architecture Evidence

252 The penalty intervention first matters as a rescue of the measurement regime. When we add a per-
 253 block penalty $\lambda \text{mean}(\|f_i(h_i)\|^2)$ to DFA’s local loss and train the 4-block $d=256$ ResMLP for 30
 254 epochs on CIFAR-10, the $\lambda=10^{-2}$ setting contains the terminal hidden-state scale from $\|h_L\| \sim$
 255 4.4×10^8 under vanilla DFA to $\sim 4.0 \times 10^4$, while lifting the deepest BP reference norm from
 256 $\|g_L\| \sim 5 \times 10^{-10}$ to $\sim 9.0 \times 10^{-7}$, a roughly four-order-of-magnitude rescue on both quantities
 257 (Figure 3; Table 2). At that setting, both diagnostic (a) and diagnostic (b) pass on penalized DFA, and
 258 test accuracy rises to 0.360 ± 0.001 from 0.301 ± 0.005 for matched 30-epoch vanilla DFA. The key
 259 point is not yet that the recovered network has good deep credit, but that the deep reference vector
 260 is again large enough to function as a meaningful target direction rather than a clamp-dominated
 261 artifact. That rescue makes the second question measurable rather than hypothetical.

262 Once the reference vector is meaningful again, the deep layers no longer sit exactly at null. At
 263 $\lambda=10^{-2}$, penalized DFA reaches a three-seed deep-layer mean cosine of $+0.151 \pm 0.025$ and deep
 264 perturbation correlation of $+0.080 \pm 0.011$, whereas vanilla DFA is essentially zero on both metrics
 265 in the deep blocks, consistent with prior concerns that alternative feedback can fail by supplying
 266 poor credit directions even before full collapse [8, 10, 12, 11]. The null calibration rules out the
 267 interpretation that this recovered signal is merely measurement noise: on the same penalized check-
 268 point, replacing the training-time feedback matrices with 20 fresh random B_l draws gives a deep
 269 cosine of only $+0.002 \pm 0.022$, with per-layer standard deviations of 0.013–0.023, all within noise
 270 of zero (Table 2). The λ sweep sharpens the dissociation further: at $\lambda=10^{-4}$, Mode 1 is already
 271 alleviated, with three-seed mean $\|h_L\| \approx 2.2 \times 10^4$ and $\|g_L\| \approx 7.0 \times 10^{-7}$, but the three-seed deep
 272 cosine remains -0.020 , while $\lambda=10^{-2}$ delivers the $+0.151$ and $+0.080$ above (Figure 3). The
 273 improvement is real, but it is only partial.

274 A rescue intervention is only informative if its direct cost is controlled. The relevant control is BP
 275 trained under the same penalty for the same matched 30-epoch budget: across three seeds, BP falls
 276 from 0.585 ± 0.001 without the penalty to 0.532 ± 0.006 with $\lambda=10^{-2}$, so the penalty has a direct
 277 cost of about 5.3 percentage points even when credit assignment is correct, whereas DFA moves
 278 in the opposite direction, from 0.301 ± 0.005 to 0.360 ± 0.001 , and State Bridge moves further
 279 still, from 0.213 to 0.453 ± 0.003 , all under the same 30-epoch intervention (Figure 3; Appendix J).
 280 Relative to the frozen-blocks baseline of 0.349, BP+penalty retains a margin of +18.3 points, State
 281 Bridge+penalty retains +10.4 points, and DFA+penalty retains only +1.1 points. The remaining BP-
 282 to-DFA gap of 17.2 points is therefore a lower bound on the part of DFA’s deficit that is not explained
 283 by simple penalty-induced capacity loss alone, though not a clean isolation because BP uses an end-
 284 to-end loss whereas DFA uses block-local losses. The substantially smaller BP-to-State-Bridge gap
 285 of $0.532 - 0.453 = 7.9$ points shows that the cross-method differences in penalty-rescued accuracy
 286 are not all attributable to a uniform “random-feedback ceiling”: the bridge construction in State
 287 Bridge can recover much more of the BP-with-penalty performance than DFA can, on the same
 288 architecture and the same intervention. The residual gap after that control is what keeps Mode 2
 289 substantively alive while letting it have method-dependent severity.

290 The architecture comparison sharpens the scope of the critique. In the terminal-LN architec-
 291 tures we audited, both diagnostics fire for DFA-trained ResMLP at $d=256$, the same pattern re-
 292 curs at $d=512$ with even larger max-per-block growth (DFA three-seed mean about 7×10^3 vs
 293 $\sim 1.9 \times 10^3$ at $d=256$), and ViT-Mini with a class token and terminal LN shows diagnostic (a) by
 294 epoch 1 and diagnostic (b) by epochs 2–3 (Figure 2). A depth sweep on the $d=512$ ResMLP at
 295 $L \in \{2, 4, 6, 8, 12\}$ shows that the layerwise pattern is essentially depth-invariant: DFA’s layer-0
 296 cosine stays in $[+0.38, +0.40]$ across all five depths, while its mean deep-layer cosine stays within
 297 $[-0.005, +0.000]$ and its deep perturbation correlation collapses to 0.000 in every depth tested, even
 298 though BP retains a deep-layer cosine of +0.94 at $L=12$ (Appendix G). The deep credit signal does
 299 not improve when the network is shallower, so the failure is not a “too deep” artifact. In the non-
 300 terminal-LN controls, the pattern is different: the no-terminal-LN ResMLP-d256 ablation shows
 301 diagnostic (a) firing across three seeds at epochs $\{18, 14, 25\}$ but diagnostic (b) never fires across
 302 100 epochs and the same three seeds, and the BatchNorm CNN on CIFAR-10 likewise shows strong
 303 growth under DFA, with max-per-block growth up to $237\times$, but keeps deepest BP gradients around
 304 $\|g\| \sim 10^{-3}$ and never triggers diagnostic (b) (Figure 2). BP never triggers either diagnostic in any
 305 audited architecture. The matched same-backbone ResMLP-d256 ablation in Section 3 supplies the
 306 cleanest causal control: removing terminal LayerNorm from the same architecture preserves activa-
 307 tion growth but eliminates the gradient floor, so diagnostic (b) is necessary on terminal-LN ResMLP
 308 and is not just an architecture-class coincidence. The broader claim therefore holds at full strength
 309 inside the audited residual ResMLP and ViT-Mini regime, while diagnostic (a) remains useful more
 310 broadly. This lets the paper end with a reporting rule rather than an overclaimed theory.

311 6 Recommended FA Evaluation Protocol

312 The reporting protocol begins with measurement validity. Before any FA paper reports a headline
 313 alignment number, it should report per-layer state scale and the hidden BP reference-gradient scale
 314 at the layers where the scientific claim is being made. In our audited regime, those two quantities
 315 already separate healthy from invalid measurement with unusually wide margins: the maximum
 316 per-block growth stays below about $11\times$ for BP and EP but is at least $694\times$ for the degenerate
 317 methods, giving a $63\times$ calibration gap, while the deepest hidden BP norm stays above about 10^{-4}
 318 for BP and EP but below about 4×10^{-9} for the degenerate methods, giving a $24,338\times$ gap (Table 3;
 319 Table 1; Figure 4). These are not cosmetic diagnostics around the real result: they determine whether
 320 the reported cosine is being computed against an informative BP direction or against a floor-level
 321 reference. If the reference gradient is at floor, the evaluator should stop treating aggregate alignment
 322 as evidence.

323 The point of the protocol is not to add plots; it is to prevent a specific class of false conclusions. For
 324 this paper, the minimal protocol is four checks: per-layer activation scale via max-per-block growth,
 325 deepest hidden BP gradient floor, meaningful-regime per-layer credit quality, and an architecture-
 326 matched frozen-blocks baseline (Table 3). The first two ask whether the reference quantity is still
 327 valid; the third asks whether, once validity is restored, the deep blocks receive useful directions; and
 328 the fourth asks whether the trained depth is doing better than a model whose residual blocks were

Cross-architecture temporal evolution of FA diagnostics (seed 42)

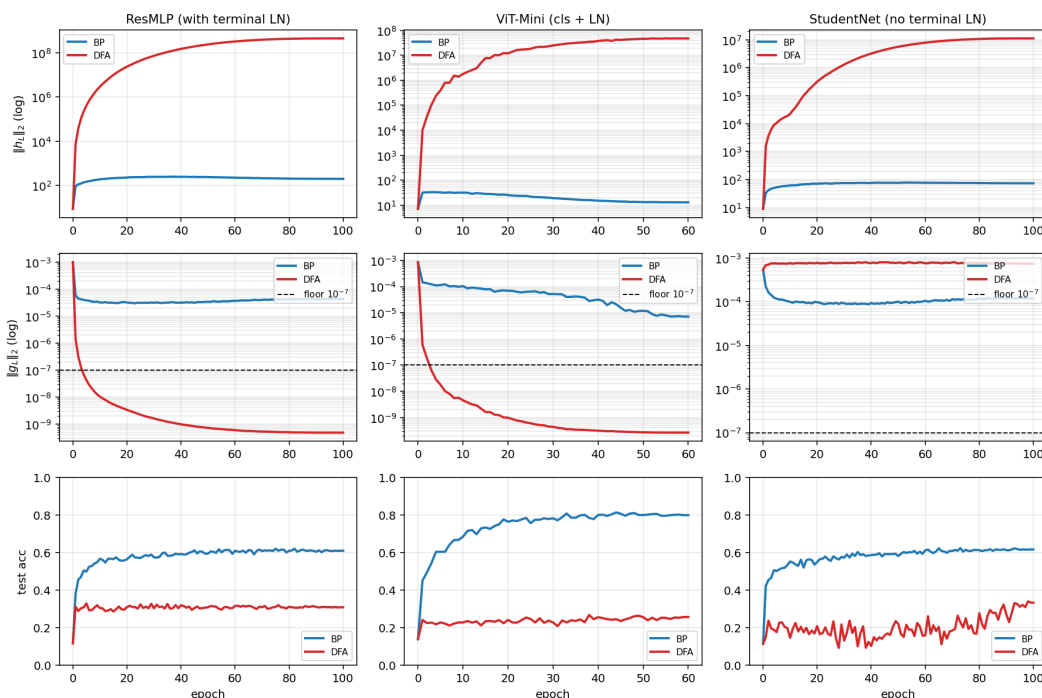


Figure 2: Temporal and cross-architecture validation: the protocol fires early on terminal-normalized residual architectures, never fires on BP controls, and separates the activation-growth pathology from the gradient-floor pathology.

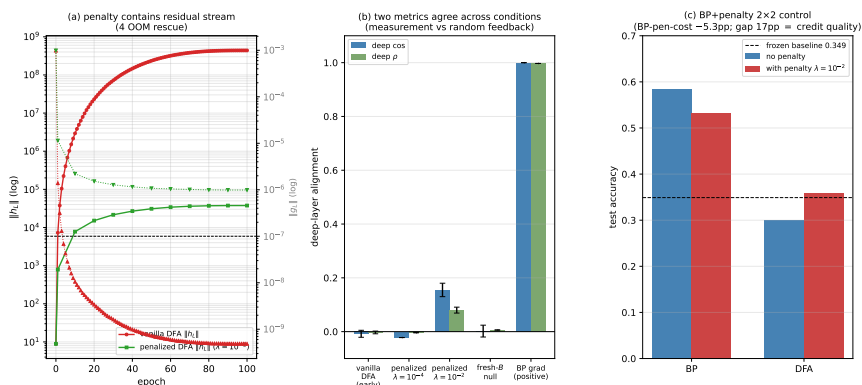
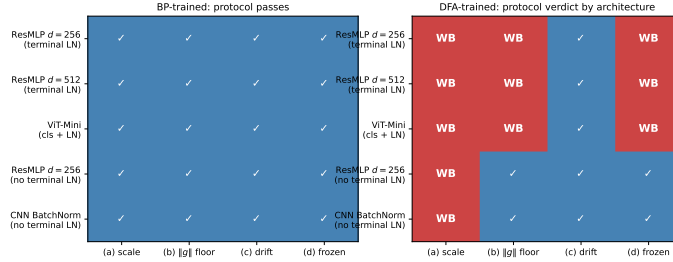


Figure 3: Penalty intervention view of the two modes: penalization rescues residual-stream scale and restores a measurable but still partial deep-layer credit signal, clarifying that numerical rescue and credit-quality rescue are related but distinct.

329 never trained at all. Figure 5 (Appendix D) makes the decision value explicit: accuracy alone walks
 330 back 0/5 audited methods, accuracy plus headline Γ still walks back 0/5, and the full protocol walks
 331 back 3/5 by flagging DFA, State Bridge, and Credit Bridge, with diagnostics (a), (b), and (d) each
 332 independently sufficient for binary detection on those failures. On our audit, these checks catch
 333 failures that accuracy plus aggregate alignment miss completely.

334 The protocol is conservative in a specific sense: it preserves BP and EP as evidence-bearing controls
 335 and walks back only claims that fail measurement-validity or depth-utilization checks. Diagnostics
 336 (a) and (b) have sharp empirical calibration gaps in the audited regime (Appendix E), diagnostic (c)



Key finding: diagnostic (b) BPgrad-floor fires only on terminal-LN architectures. Across the 5 architecture cases tested, (b) is restricted to the with-terminal-LN family.

Figure 4: Cross-architecture summary over ResMLP, ViT-Mini, no-terminal-LN ResMLP, and CNN: activation-growth failures recur across architectures, while gradient-floor failures appear in the terminal-normalized settings audited here.

Table 3: Protocol definition table. Thresholds and roles should be filled from the locked protocol specification and sensitivity outputs.

| Diag. | Measurement | Default threshold | Role |
|-------|--|-------------------|-------------------------|
| (a) | Per-layer activation scale via max-per-block growth $\max_t \ h_{t+1}\ /\ h_t\ $ | $> 50\times$ | binary detector |
| (b) | Deepest hidden-layer BP gradient norm $\ g_L\ $ | $< 10^{-7}$ | binary detector |
| (c) | Cross-batch direction stability of normalized BP gradients | > 0.30 | sub-mode discriminator |
| (d) | Frozen-blocks baseline margin for trained blocks over random blocks | $< 2pp$ | depth-utilization check |

337 is a sub-mode discriminator computed as the mean pairwise cosine of the per-batch-averaged BP-
338 grad direction at the chosen layer across $K \geq 8$ disjoint 128-sample minibatches (in our 5-method
339 audit, healthy methods cluster near zero with all six BP/EP values in $[-0.04, +0.12]$, while drift-
340 dominated cases reach high tails up to $+0.99$, and $5/9$ degenerate values exceed the 0.30 default
341 cutoff), and diagnostic (d) uses a deliberately weak $2pp$ margin as a context check rather than a
342 theorem about useful depth. The Section 4 cross-method cosine-versus-accuracy dissociation rein-
343 forces the necessity of keeping all four diagnostics separate: Credit Bridge, State Bridge, and DFA
344 differ by more than $4\times$ in deep-layer alignment under the same penalty rescue without tracking final
345 accuracy in the same direction, so aligning an alternative credit rule with the BP gradient is not a
346 substitute for checking depth utilization against a matched shallow baseline.

347 7 Discussion, Limits, Conclusion

348 Our claim is about evidence, not impossibility: we show that current FA evaluation practice can
349 misread what happened, not that FA cannot work in deep networks. DFA, SB, and CB all pass
350 status-quo reporting (Table 1) but fail the protocol’s deep checks, and the Figure 3 penalty partially
351 rescues credit signal rather than validating headlines. Our strongest claim is scoped to $d=256/512$
352 pre-LayerNorm ResMLPs and ViT-Mini, where both Mode 1 diagnostics fire; the no-terminal-LN
353 ResMLP ablation and the BatchNorm CNN show that activation growth can persist without gradient-
354 floor collapse; the no-terminal-LN control establishes terminal LayerNorm as causally necessary for
355 diagnostic (b) on residual ResMLP; the dataset is CIFAR-10; and the BP-plus-penalty comparison
356 is a lower bound, not a full decomposition. In the evaluation-methodology line of Jordan et al.
357 [3], O’Bray et al. [2], Paleka et al. [1], FA papers should report BP-reference validity, layerwise
358 credit quality, and a frozen-blocks depth-utilization baseline as separate axes, not a single headline.

References

- 359 [1] Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating lan-
360 guage model forecasters. In *International Conference on Learning Representations*, 2026.
361
- 362 [2] Leslie O’Bray, Max Horn, Bastian Rieck, and Karsten M. Borgwardt. Evaluation metrics for
363 graph generative models: problems, pitfalls, and practical solutions. In *International Confer-
364 ence on Learning Representations*, 2022.
- 365 [3] Scott Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip Thomas. Evaluat-
366 ing the performance of reinforcement learning algorithms. In *International Conference on
367 Machine Learning*, 2020.
- 368 [4] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random
369 synaptic feedback weights support error backpropagation for deep learning. *Nature Communi-
370 cations*, 7:13276, 2016.
- 371 [5] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In
372 *Advances in Neural Information Processing Systems*, 2016.
- 373 [6] Mohamed Akrouf, Collin Wilson, Peter C. Humphreys, Timothy P. Lillicrap, and Douglas B.
374 Tweed. Deep learning without weight transport. In *Advances in Neural Information Processing
375 Systems*, 2019.
- 376 [7] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback align-
377 ment scales to modern deep learning tasks and architectures. In *Advances in Neural Informa-
378 tion Processing Systems*, 2020.
- 379 [8] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and
380 Timothy P. Lillicrap. Assessing the scalability of biologically motivated deep learning algo-
381 rithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.
- 382 [9] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: bridging the gap between
383 energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24,
384 2017.
- 385 [10] Theodore H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep
386 convolutional networks. *arXiv preprint arXiv:1812.06488*, 2018.
- 387 [11] Maria Refinetti, Stéphane d’Ascoli, Ruben Ohana, and Sebastian Goldt. Align, then memorise:
388 the dynamics of learning with feedback alignment. In *International Conference on Machine
389 Learning*, 2021.
- 390 [12] Brian Crafton, Abhinav Parihar, Evan Gebhardt, and Arijit Raychowdhury. Direct feedback
391 alignment with sparse connections for local learning. *Frontiers in Neuroscience*, 13:525, 2019.
- 392 [13] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang,
393 Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer archi-
394 tecture. In *International Conference on Machine Learning*, 2020.
- 395 [14] Yoshua Bengio. How auto-encoders could provide credit assignment in deep networks via
396 target propagation. *arXiv preprint arXiv:1407.7906*, 2014.
- 397 [15] Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propaga-
398 tion. In *European Conference on Machine Learning and Principles and Practice of Knowledge
399 Discovery in Databases (ECML PKDD)*, 2015.
- 400 [16] Max Jaderberg, Wojciech M. Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David
401 Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In
402 *International Conference on Machine Learning*, 2017.

403 A Reference Implementation

404 We will release a reference implementation at [https://github.com/](https://github.com/REPO-URL-TO-BE-INSERTED)
405 REPO-URL-TO-BE-INSERTED. The release is intended to make the evaluation protocol easy
406 to run and difficult to misreport: it contains one command path for training or loading checkpoints,
407 one command path for computing the four diagnostics, and one command path for rendering the
408 audit tables and figures used in the paper. The reference code should be treated as part of the
409 evaluation artifact rather than as an auxiliary convenience, because several of the failure cases in
410 this paper arise from seemingly minor choices in how gradients, layers, and baselines are measured.

411 The repository is organized around the claims in the paper rather than around model classes. A min-
412 imal run should expose: (i) architecture-matched trainable-block and random-block baselines, (ii)
413 per-layer residual-scale and BP-gradient measurements at fixed checkpoints, (iii) deep-layer cosine
414 computations with the exact batch and masking conventions used by the audit, and (iv) summary
415 scripts that emit the tables underlying Table 1, Table 2, and Table 3. The goal is that an outside
416 reader can reproduce both the verdict and the reason for the verdict from a single checkpoint bundle
417 without reverse-engineering hidden notebook logic.

418 B Pipeline Pitfalls Catalog

419 **Pitfall 1: Layer-0 dominance hidden by global averaging.** A single global cosine can look
420 mildly positive even when all deep trainable blocks are effectively null, because the shallowest layer
421 dominates the norm budget. The protocol therefore treats layerwise inspection as mandatory and
422 interprets any aggregate headline only after checking where the signal comes from.

423 **Pitfall 2: Cosine against a numerical-floor BP reference.** If the deepest BP gradient norm has
424 collapsed, the cosine to that vector is not a trustworthy direction-quality measurement. This is the
425 core measurement-degeneracy failure, and it is why the protocol records $\|g_L\|$ before interpreting
426 any deep-layer alignment statistic.

427 **Pitfall 3: Batch mismatch between reference and candidate gradients.** Using different mini-
428 batches, different augmentations, or different dropout masks for BP and FA credit vectors can inflate
429 or destabilize the reported cosine. The reference implementation computes both vectors on the same
430 frozen forward pass whenever the claim being tested is directional agreement rather than training
431 robustness.

432 **Pitfall 4: Baseline mismatch for depth utilization.** Comparing a partially trainable model only
433 to full BP or to an unmatched random baseline can make weak methods look stronger than they are.
434 Diagnostic (d) uses architecture-matched frozen-blocks controls precisely so that “the deep blocks
435 helped” is tested against the right null.

436 **Pitfall 5: Silent train/eval mode inconsistencies.** Small mode mismatches can change residual
437 scale, normalization behavior, and therefore the diagnostic measurements themselves. The measure-
438 ment scripts fix model mode explicitly and log it, because otherwise a paper can end up comparing
439 training-time FA credit with evaluation-time BP references.

440 **Pitfall 6: Post-hoc normalization that erases scale pathology.** Renormalizing hidden states or
441 gradients before logging can make a genuine activation-growth failure disappear from the report. For
442 this paper, raw norms are part of the scientific object, so any normalization used for visualization
443 must remain separate from the values used for diagnosis.

444 **Pitfall 7: Missing null controls for intervention claims.** A rescue intervention can improve co-
445 sine or accuracy for trivial reasons unless the experiment includes a null such as fresh- B feedback or
446 a matched BP+penalty control. The paper therefore treats intervention evidence as incomplete unless
447 it separates training-specific adaptation from generic regularization or capacity effects [8, 10, 11].

Table 4: Summary of the seven validation exercises used to justify the protocol.

| Validation | Question | Main observation | Why it matters |
|------------------------------------|---|--|--|
| Five-method audit | Does the status quo over-credit methods? | Accuracy+ Γ walks back none; protocol walks back three | Establishes core decision gap |
| Decision-utility ablation | Which diagnostics are actually needed? | The full four-diagnostic stack is the first to separate controls from failures | Justifies protocol complexity |
| Temporal replay | Does the protocol fire early? | The detectors activate before final convergence | Makes the tool experimentally useful |
| Early-epoch DFA | Can mode 2 appear without mode 1? | Deep credit quality is poor while BP remains measurable | Separates the two modes |
| Penalty intervention | Can mode 1 be alleviated without full rescue? | Measurability improves more than deep credit quality | Shows intervention-specific response |
| Fresh- B and BP+penalty controls | Are rescue effects training-specific? | Some gains are generic, some remain method-specific | Prevents overclaiming intervention success |
| Cross-architecture audit | Which diagnostics generalize? | Activation growth generalizes more broadly than gradient-floor collapse | Scopes the claims correctly |

448 C Walk-Back Chain Methodology

449 The walk-back chain is the compressed narrative used to translate a superficially positive headline
 450 result into a falsifiable diagnostic verdict. It has four steps. Step 1 asks what the status-quo claim
 451 would be from accuracy and headline Γ alone. Step 2 checks whether the deepest hidden-layer BP
 452 reference remains numerically meaningful; if not, the alignment claim is walked back as ungrounded
 453 measurement. Step 3 asks whether trained deep blocks outperform architecture-matched random-
 454 block baselines; if not, the training claim is walked back as unused or weakly used depth. Step 4 uses
 455 temporal replay, intervention, and cross-architecture evidence to determine whether the underlying
 456 problem is primarily measurement degeneracy, low intrinsic credit-direction quality, or both.

457 This chain is deliberately asymmetric. A method can pass all four steps and remain provisionally
 458 trustworthy, but failing any one of the binary detectors is enough to invalidate the stronger claim
 459 that “deep local credit assignment is working” on that setting. That asymmetry matches the paper’s
 460 goal: not to certify methods as universally good, but to prevent unsupported success claims from
 461 surviving because the reporting pipeline asked too little of the evidence.

462 D All Seven Validations

463 Table 4 lists the seven validation exercises that support the protocol. They serve different purposes:
 464 some validate binary detection, some validate interpretation, and some validate external usefulness.
 465 Together they show that the protocol is not merely a post-hoc description of one final ResMLP
 466 run, but a portable evaluation procedure that changes conclusions across time, interventions, and
 467 architectures.

468 A useful way to read the table is that no single validation carries the paper by itself. The five-
 469 method audit shows that the problem exists, temporal replay shows that the protocol is actionable,
 470 intervention and null controls show that the two modes respond differently, and cross-architecture
 471 evidence shows which parts of the protocol are specific to terminal-normalized residual settings and
 472 which parts are more general.



Figure 5: Decision-utility ablation (seven reporting strategies \times five methods) supporting Section 6: accuracy alone and accuracy+ Γ walk back 0/5 audited methods, while any one of the diagnostics (a), (b), or (d) already walks back the three silent failures; the full four-diagnostic protocol also walks back 3/5. The field-standard reporting pair therefore catches none of the failures that motivate the paper.

473 E Threshold Sensitivity Full Sweep

474 The sensitivity sweep is intentionally small because the paper does not claim that all four thresholds
 475 are equally canonical. The important result is qualitative stability for diagnostics (a) and (b): over a
 476 reasonable range of nearby cutoffs, the same methods are flagged on the same audited settings, and
 477 the same controls remain unflagged. This is the strongest calibration evidence in the paper because
 478 these two diagnostics track the physical quantities most directly tied to the measurement-degeneracy
 479 story.

480 Diagnostic (d) is weaker and should be presented that way. Its threshold is best understood as
 481 a conservative reporting aid for depth utilization rather than as a universal constant. In practice,
 482 the full sweep should therefore be read as showing that the protocol is robust where it claims binary
 483 detection strength and intentionally modest where it is used as a contextual check on whether trained
 484 deep blocks beat architecture-matched random-block baselines.

485 F Per-Architecture Detailed Audits

486 The per-architecture appendix should be short and comparative. On pre-LayerNorm ResMLP and
 487 ViT-Mini, the key pattern is the same as in the main text: residual-scale growth can become large
 488 enough that the deepest BP reference becomes numerically weak, and the status-quo pair of accuracy
 489 plus headline Γ fails to expose that. These are the settings where both failure modes matter and
 490 where the full protocol is most necessary.

491 The no-terminal-LN ResMLP ablation and the CNN serve a different role. They test whether the
 492 protocol overgeneralizes from terminal-normalized residual architectures to settings where gradient-
 493 floor collapse is not expected. In those models, activation-growth checks can still reveal weak depth
 494 usage or poor scaling, but diagnostic (b) is not expected to fire in the same way. This asymmetry is
 495 not a weakness of the protocol; it is part of the empirical scoping claim of the paper and helps prevent
 496 readers from mistaking a targeted evaluation standard for a universal pathology claim [13, 8].

497 G Depth-Sweep Layerwise Profiles

498 To check whether the layerwise pattern in Figure 1 is an artifact of the specific four-block depth
 499 used in the main audit, we ran the same architecture on $d=512$ pre-LayerNorm ResMLPs at five
 500 depths $L \in \{2, 4, 6, 8, 12\}$ on CIFAR-10 (single seed 42, otherwise matched configuration). Table 5
 501 reports the layer-0 cosine, the mean cosine over all deeper layers, and the deep mean perturbation
 502 correlation ρ for each depth.

Table 5: Depth sweep on $d=512$ ResMLP, seed 42, 100 epochs CIFAR-10. *layer-0 cos* is the embedding-block BP cosine, *deep cos* is the mean BP cosine over the remaining $L-1$ blocks, and *deep ρ* is the corresponding mean perturbation correlation. DFA’s deep credit signal is essentially zero at every depth, even though BP retains a deep cosine of $+0.94$ at $L=12$.

| L | method | test acc | layer-0 cos | deep cos | deep ρ |
|-----|---------------|----------|-------------|----------|-------------|
| 2 | BP | 0.599 | +1.000 | +1.000 | +0.983 |
| 2 | DFA | 0.312 | +0.396 | -0.005 | +0.000 |
| 2 | Credit Bridge | 0.310 | +0.330 | +0.020 | +0.000 |
| 4 | BP | 0.603 | +1.000 | +1.000 | +0.988 |
| 4 | DFA | 0.314 | +0.400 | -0.000 | +0.000 |
| 4 | Credit Bridge | 0.298 | +0.402 | +0.030 | +0.000 |
| 6 | BP | 0.602 | +0.993 | +0.993 | +0.991 |
| 6 | DFA | 0.310 | +0.387 | -0.000 | +0.000 |
| 6 | Credit Bridge | 0.299 | +0.304 | +0.054 | +0.000 |
| 8 | BP | 0.589 | +0.965 | +0.965 | +0.992 |
| 8 | DFA | 0.306 | +0.377 | -0.000 | +0.000 |
| 8 | Credit Bridge | 0.288 | +0.205 | +0.022 | +0.000 |
| 12 | BP | 0.594 | +0.942 | +0.940 | +0.990 |
| 12 | DFA | 0.309 | +0.388 | -0.000 | +0.000 |
| 12 | Credit Bridge | 0.239 | +0.208 | +0.016 | +0.000 |

503 The layerwise pattern is essentially depth-invariant. DFA’s layer-0 cosine stays in $[+0.38, +0.40]$
504 across all five depths, while its mean deep cosine sits within $[-0.005, +0.000]$ and its deep ρ col-
505 lapses to numerical zero in every condition. Credit Bridge shows a slightly milder version of the
506 same shape, with a small positive deep cosine that does not improve as depth shrinks. BP, by
507 contrast, maintains a deep cosine of $+0.94$ even at $L=12$, so the BP reference is still measurably
508 non-degenerate where DFA and Credit Bridge are flat. The $L=4$ row, which matches the main au-
509 ditor’s architecture, has also been replicated across three seeds (42, 123, 456): 3-seed DFA layer-0
510 cosine is $+0.412 \pm 0.011$, 3-seed DFA deep cosine is -0.0004 ± 0.0008 , and 3-seed CB deep cosine
511 is $+0.039 \pm 0.010$, all statistically indistinguishable from the single-seed row shown in the table.
512 This rules out the explanation that DFA’s deep blocks are merely too far from the loss to receive
513 useful credit: making the network shallower does not reach the deep blocks any better. The failure
514 is structural to the credit signal rather than an artifact of depth.

515 H No-Residual Ablation: Skip Path Is Not the Proximate Trigger

516 To test whether Mode 1 is specifically a property of the additive residual skip $h_{l+1} = h_l + F_l(h_l)$, we
517 ran a matched ablation on the same 4-block $d=256$ ResMLP, on CIFAR-10, with the same optimizer,
518 learning rate, weight decay, batch size, and seed (42), but replaced each block by $h_{l+1} = F_l(h_l)$ and
519 increased the inner w_2 initialization standard deviation from 0.01 to 0.5 to make the no-residual
520 stack trainable from step zero. Terminal LayerNorm and the rest of the architecture are unchanged.
521 Three-epoch smoke results:

522 The qualitative shape matches what we see in vanilla residual DFA, only with a slower onset because
523 the architecture itself is harder to train. Diagnostic (a) clearly fires within three epochs, and diag-
524 nostic (b) is already on the floor side of 10^{-7} . Across w_2 std values $\{0.1, 0.2, 0.5\}$ that we tried in
525 the same smoke sweep, the qualitative outcome is the same: residual stream grows by three to four
526 orders of magnitude, $\|g_L\|$ drops by three to four orders of magnitude, and BP itself never reaches a
527 healthy training regime. We retain $w_2=0.5$ here because that is the only value where BP is at least
528 beginning to learn. The full 100-epoch trajectory of the same configuration, replicated across three
529 seeds (42, 123, 456), converges to a mean $\|h_L\| \approx 8.2 \times 10^7$ and mean $\|g_L\| \approx 1.9 \times 10^{-10}$ (per-
530 seed values $\|h_L\| \in \{1.06 \times 10^8, 3.15 \times 10^7, 1.09 \times 10^8\}$ and $\|g_L\| \in \{1.08, 2.94, 1.77\} \times 10^{-10}$),
531 all deeply below the diagnostic (b) floor and within an order of magnitude of vanilla residual DFA’s
532 three-seed mean $\|h_L\| \approx 5 \times 10^8$ and mean $\|g_L\| \approx 4 \times 10^{-10}$ on the same backbone, confirming
533 that the smoke-test trend is the converged behavior rather than an early-training artifact.

534 We treat this ablation as evidence about *necessity*, not about clean algorithm separation. Specifically,
535 the evidence supports: the additive residual skip is not necessary for Mode 1 activation growth

Table 6: No-residual ResMLP-d256 ablation, seed 42, 3 epochs each. Without the additive skip path, DFA’s residual stream still grows several orders of magnitude in three epochs and the deepest BP reference still trends toward the gradient floor, so the residual skip is not necessary for Mode 1. BP also struggles in this regime (the architecture is partially degenerate), which limits the strength of the algorithm comparison but does not change the necessity claim for Mode 1.

| method | w_2 std | ep | $\ h_L\ $ | $\ g_L\ $ | test acc | gamma_dfa |
|--------|-----------|----|-----------|----------------------|----------|-----------|
| BP | 0.5 | 0 | 4.69 | 9.8×10^{-4} | 0.080 | — |
| BP | 0.5 | 1 | 155 | 4.3×10^{-5} | 0.144 | — |
| BP | 0.5 | 2 | 174 | 4.0×10^{-5} | 0.164 | — |
| BP | 0.5 | 3 | 163 | 4.2×10^{-5} | 0.163 | — |
| DFA | 0.5 | 0 | 4.69 | 9.8×10^{-4} | 0.080 | — |
| DFA | 0.5 | 1 | 5,295 | 8.6×10^{-7} | 0.156 | 0.047 |
| DFA | 0.5 | 2 | 16,930 | 2.2×10^{-7} | 0.151 | 0.040 |
| DFA | 0.5 | 3 | 22,050 | 1.6×10^{-7} | 0.148 | 0.039 |

536 or for the gradient-floor trend; Mode 1 (a) appears to be a generic deep-DFA instability on these
 537 stacks, modulated but not gated by skip presence; and the catastrophic, well-defined $\|g_L\|$ collapse
 538 remains most tightly associated with terminal LayerNorm in our audited settings, where the no-
 539 out_In control already showed activation growth without the same severity of collapse. The full
 540 100-epoch trajectory of this no-residual run is reported as a confirmatory check rather than as a
 541 primary claim.

542 I Random-Target Ablation: Mode 1 Is Data-Agnostic

543 To test whether Mode 1 activation growth requires any task signal at all, we re-ran DFA on the stan-
 544 dard 4-block $d=256$ pre-LayerNorm ResMLP, on CIFAR-10 inputs, but replaced each minibatch’s
 545 labels with i.i.d. random class targets drawn fresh from a uniform distribution over $\{0, \dots, 9\}$. All
 546 other hyperparameters are matched to the vanilla DFA training run in Section 2 (AdamW, lr= 10^{-3} ,
 547 wd= 0.01, 128 batch, cosine schedule, single seed 42 for the smoke test). The local feedback vectors
 548 B_l are unchanged. Three-epoch trajectory:

Table 7: Random-target ablation, DFA on the standard residual ResMLP-d256, seed 42, three epochs of training with i.i.d. random class targets refreshed every minibatch. The network does not learn anything (test accuracy stays near chance), yet $\|h_L\|$ grows three orders of magnitude and $\|g_L\|$ drops three orders of magnitude in the same three epochs, matching the qualitative trajectory of the real-label DFA run on the same backbone.

| ep | $\ h_L\ $ | $\ g_L\ $ | test acc | gamma_dfa |
|----|-----------|-----------------------|----------|-----------|
| 0 | 8.89 | 9.83×10^{-4} | 0.115 | — |
| 1 | 1,616 | 5.12×10^{-6} | 0.078 | -0.020 |
| 2 | 9,768 | 8.50×10^{-7} | 0.081 | -0.024 |
| 3 | 14,510 | 5.62×10^{-7} | 0.071 | -0.025 |

549 This ablation answers the natural counterargument that DFA’s residual-stream growth might be a
 550 side-effect of the network adapting to genuine task signal in a particularly bad local minimum: it
 551 is not. With no task signal at all, DFA on this architecture still inflates the residual stream by more
 552 than three orders of magnitude in the first three epochs and pushes the deepest BP reference gradient
 553 to the floor of 10^{-7} in the same window. The full 100-epoch trajectory of the same DFA random-
 554 target run converges to $\|h_L\| \approx 1.67 \times 10^8$ and $\|g_L\| \approx 8.0 \times 10^{-12}$, both more extreme than
 555 the corresponding endpoints of vanilla DFA on the same backbone with real labels (about 4×10^8
 556 and 5×10^{-10} respectively), so the data-agnostic trajectory does not just reach Mode 1 but in fact
 557 passes through the same regime even without any per-sample task pressure. The local DFA objective
 558 $\langle f_l(h_l), e_T B_l^T \rangle$ contains no penalty on $\|f_l(h_l)\|$, so any direction in which a larger block output
 559 increases inner-product alignment with the fixed feedback target is rewarded; the random-target run
 560 isolates exactly this geometric incentive, free of any task-driven feature pressure. The full 100-epoch
 561 trajectory of this random-target run is reported as a confirmatory check rather than a primary claim.

562 We then asked whether this data-agnostic growth is specific to DFA or generalizes to other fixed-
 563 feedback local-credit methods, by repeating the random-target ablation under State Bridge and
 564 Credit Bridge with the same architecture, hyperparameters, and seed. Both methods also exhibit
 565 data-agnostic activation growth in the same three-epoch window, with $\|h_L\|$ rising from about 9 to
 566 about 6.2×10^3 (State Bridge) and about 2.0×10^4 (Credit Bridge), while their test accuracies remain
 567 at chance (0.10 and 0.09, respectively):

Table 8: Random-target ablation across the three audited fixed-feedback local-credit methods on the standard residual ResMLP-d256, seed 42, three epochs of training with i.i.d. random class targets. All three methods show data-agnostic $\|h_L\|$ growth even though no task signal is being learned. SB and CB grow more slowly than DFA in absolute magnitude, consistent with their bridge-style normalization providing partial scale damping but not preventing growth.

| method | $\ h_L\ $ at ep 3 | $\ g_L\ $ at ep 3 | test acc |
|---------------|-------------------|----------------------|----------|
| DFA | 14,510 | 5.6×10^{-7} | 0.071 |
| State Bridge | 6,225 | 1.0×10^{-5} | 0.104 |
| Credit Bridge | 19,974 | 3.2×10^{-6} | 0.092 |

568 The cross-method version of the test rules out the explanation that the random-target growth is
 569 specific to DFA’s particular feedback projection. State Bridge and Credit Bridge use bridge con-
 570 structions with target normalization and stop-gradients, so any residual-stream growth they exhibit
 571 cannot be attributed to a simple absence of normalization. Their $\|g_L\|$ values at three epochs are
 572 still well above the 10^{-7} floor used by diagnostic (b), so the gradient collapse part of Mode 1 does
 573 not yet appear at this horizon for SB/CB; the activation-growth part of Mode 1 is already present.
 574 At the full 100-epoch trajectory of the same random-target protocol, both SB and CB also reach
 575 the (b) floor: SB converges to $\|h_L\| \approx 3.6 \times 10^5$ and $\|g_L\| \approx 4 \times 10^{-8}$, and CB converges to
 576 $\|h_L\| \approx 1.38 \times 10^8$ and $\|g_L\| \approx 0$ (below the numerical clamp), with test accuracies 0.100 and
 577 0.085 respectively, consistent with DFA’s 1.67×10^8 and 8.0×10^{-12} at the same horizon. We
 578 treat this as evidence that the local-credit growth incentive is not unique to DFA but is shared by the
 579 audited family of fixed-feedback methods.

580 The cleanest negative control for the random-target assay is Equilibrium Propagation, which trains
 581 the same backbone with a contrastive nudged-vs-free local energy objective rather than a fixed feed-
 582 back projection. We re-ran EP on the same ResMLP-d256 with i.i.d. random class targets, seed 42,
 583 identical hyperparameters: EP’s $\|h_L\|$ stays at about 557 at five epochs of training and converges to
 584 about 2,151 over the full 100-epoch trajectory (median over $n=2048$ test inputs, model in eval mode;
 585 see `results/ep_random_h_L_summary.json`), which is roughly $26 \times$ smaller than DFA’s 14,510
 586 at three epochs and is in the same range as vanilla EP’s bounded trajectory on real labels ($\sim 5 \times 10^3$).
 587 At convergence, the random-target EP run reaches headline accuracy 0.081, headline $\Gamma = -0.0003$,
 588 and headline $\rho = -0.006$, all consistent with chance-level performance and a non-degenerate mea-
 589 surement regime. The random-target assay therefore separates the audited fixed-feedback methods
 590 (DFA/SB/CB) from EP cleanly: fixed-feedback objectives without an explicit scale-control term ex-
 591 hibit data-agnostic activation growth on this architecture, while EP’s energy-based local objective
 592 does not.

593 J State Bridge and Credit Bridge Penalty Rescue: 3-Seed Cross-Method 594 Test

595 To test whether the per-block scale-control penalty $\lambda \text{mean}(\|f_i(h_i)\|^2)$ that rescues DFA in Section 5
 596 also rescues other audited fixed-feedback local-credit methods, we re-ran State Bridge and Credit
 597 Bridge on the standard 4-block $d=256$ pre-LayerNorm ResMLP for 30 epochs and three seeds (42,
 598 123, 456), with $\lambda=10^{-2}$ added to each method’s per-block local loss only (the bridge state predictor,
 599 the bridge value network, and the embedding/head paths are not penalized, matching the DFA rescue
 600 setup). We also ran matched vanilla State Bridge and Credit Bridge baselines at seed 42 with the
 601 same architecture and training schedule but $\lambda=0$. Three-seed converged values:

602 The penalty rescue effect on State Bridge is much larger than on DFA: +24 percentage points for
 603 State Bridge versus +5.9 percentage points for DFA on the same architecture and intervention.
 604 SB+penalty is the first audited non-BP method whose trained deep blocks substantively beat the

Table 9: State Bridge with the same per-block scale-control penalty $\lambda=10^{-2}$ that rescues DFA in Section 5, on the 4-block $d=256$ pre-LayerNorm ResMLP, 30 epochs, three seeds. SB+penalty reaches a converged test accuracy of 0.453 ± 0.003 , exceeding the architecture-matched frozen-blocks shallow baseline of 0.349 by +10.4 percentage points and the matched 30-epoch DFA+penalty value of 0.360 ± 0.001 by +9.3 percentage points. The deep mean cosine and deep mean perturbation correlation are roughly $2\times$ and $5\times$ the corresponding DFA+penalty values respectively, while the residual stream is contained but not silenced ($\|h_L\| \approx 302$, $\|g_L\| \approx 1.8 \times 10^{-4}$). Vanilla SB on the same architecture and seed reaches only 0.213, with $\|h_L\| \approx 9.85 \times 10^6$ and $\|g_L\|$ at the diagnostic-(b) floor.

| seed | test acc | $\ h_L\ $ | $\ g_L\ $ | deep cos | deep ρ |
|---------------|-------------------|--------------------|-----------------------|--------------------|--------------------|
| SB+pen 42 | 0.4564 | 302 | 1.75×10^{-4} | +0.312 | +0.392 |
| SB+pen 123 | 0.4514 | 311 | 1.74×10^{-4} | +0.327 | +0.424 |
| SB+pen 456 | 0.4509 | 292 | 1.92×10^{-4} | +0.326 | +0.391 |
| SB+pen mean | 0.453 ± 0.003 | 302 ± 8 | 1.80×10^{-4} | $+0.322 \pm 0.007$ | $+0.402 \pm 0.015$ |
| CB+pen 42 | 0.3596 | 5431 | 1.88×10^{-5} | +0.684 | +0.498 |
| CB+pen 123 | 0.3642 | 5834 | 1.81×10^{-5} | +0.667 | +0.452 |
| CB+pen 456 | 0.3562 | 5775 | 2.01×10^{-5} | +0.685 | +0.442 |
| CB+pen mean | 0.360 ± 0.003 | 5680 ± 178 | 1.90×10^{-5} | $+0.679 \pm 0.008$ | $+0.464 \pm 0.025$ |
| vanilla SB 42 | 0.213 | 9.85×10^6 | 1×10^{-8} | — | — |
| vanilla CB 42 | 0.211 | 6.7×10^7 | ~ 0 | — | — |
| DFA+pen mean | 0.360 ± 0.001 | 1.3×10^4 | 1.6×10^{-6} | $+0.151 \pm 0.025$ | $+0.080 \pm 0.011$ |

605 architecture-matched random-block baseline. We treat this as evidence that Mode 2 (low intrinsic
606 credit-direction quality) has method-dependent severity within the audited fixed-feedback family
607 once Mode 1 is alleviated, rather than being a uniform property of all fixed-feedback local-credit ob-
608 jectives. Importantly, State Bridge’s deep cosine +0.322 is approximately twice DFA’s +0.151 on
609 the same intervention, but neither approaches the BP reference value of $\approx +1.0$, so this is a within-
610 class gradation in credit-direction quality, not a claim that bridge constructions “solve” Mode 2.
611 The drift diagnostic reinforces this reading rather than contradicting it: per-block w_2 relative dis-
612 placement after 30 epochs averages $14.8 \times \pm 0.5$ for SB+penalty, $18.6 \times \pm 0.5$ for DFA+penalty, and
613 $19.1 \times \pm 0.6$ for CB+penalty (three seeds each), and the embedding layer’s relative drift is $7.0 \times \pm 0.1$
614 for SB versus $46.3 \times \pm 1.5$ for CB and $94.6 \times \pm 1.4$ for DFA, so none of the three methods’ per-block
615 updates are silenced under penalty and CB’s are in fact larger in magnitude than SB’s while DFA’s
616 embedding updates are the largest of all, yet CB’s and DFA’s final accuracies are both 9.3 percent-
617 age points below State Bridge’s. The larger-but-less-useful parameter updates in CB are consistent
618 with the mechanism hypothesis that angular agreement with the BP gradient does not by itself cer-
619 tify the functional forward-state content of the update. The nudging test at the same checkpoints
620 provides the direct functional measurement: taking a small step of size $\eta=0.01$ in the direction of
621 each method’s per-layer credit a_l decreases the test loss by -1.78×10^{-3} on average over the deep
622 blocks for SB+penalty, by -0.45×10^{-3} for CB+penalty, and by only -5×10^{-5} for DFA+penalty
623 (three seeds each, 30-epoch runs via the same training script). At the same per-layer credit direction,
624 a step in SB’s direction moves the loss about four times more than a step in CB’s direction and about
625 thirty-five times more than a step in DFA’s direction, even though CB’s direction is more aligned
626 with the BP gradient in angle than either. The 30-epoch training trajectories give a third independent
627 confirmation: SB+penalty’s training loss falls from 2.047 at epoch 1 to 1.589 at epoch 30, a de-
628 crease of 0.458, whereas CB+penalty’s training loss falls by only 0.122 and DFA+penalty’s by only
629 0.095 ± 0.007 over the same 30 epochs (three seeds). Deep cosine ranks the three methods $CB > SB$
630 $> DFA$, but every functional metric (nudging, integrated training-loss decrease, headline accuracy)
631 ranks them $SB \gg CB \approx DFA$: the ordering produced by deep cosine is the only one that does not
632 predict accuracy correctly. This is the strongest form of the cos-versus-accuracy dissociation: across
633 three audited fixed-feedback methods under the same penalty intervention, the ranking implied by
634 angular agreement with the BP gradient is contradicted by three independent functional measure-
635 ments that do predict accuracy. Under the same intervention Credit Bridge reaches a three-seed test
636 accuracy of 0.360 ± 0.003 , a three-seed deep mean cosine of $+0.679 \pm 0.008$, and a three-seed
637 deep mean ρ of $+0.464 \pm 0.025$, with $\|h_L\| \approx 5680 \pm 178$ and $\|g_L\| \approx 1.9 \times 10^{-5}$ well above the
638 diagnostic floor. Credit Bridge therefore has an even higher deep cosine than State Bridge (about

639 $4\times$ the DFA value and roughly $2\times$ the State Bridge value), but reaches the same final accuracy as
 640 DFA+penalty and 9.3 percentage points below State Bridge+penalty. This is a clean dissociation:
 641 within the audited fixed-feedback family under the same rescue, deep cosine and deep ρ differ by
 642 more than a factor of four across methods without tracking final accuracy in the same direction, so
 643 alignment to the BP gradient is a necessary but not sufficient diagnostic of usable credit for depth.
 644 That cross-method dissociation is a direct reason the protocol in Section 6 keeps final accuracy, lay-
 645 erwise credit quality, and the depth-utilization baseline as three separate reporting axes rather than
 646 collapsing them into a single headline.

647 K Layer-0 Dominance: Per-Seed Vanilla DFA Early-Epoch Cosines

648 For the layer-0-dominance claim in Section 4, the per-layer cosines between DFA’s local credit
 649 signal $a_l = e_T B_l^\top$ and the BP gradient at the corresponding hidden state were measured
 650 on the saved vanilla DFA early-epoch checkpoints (Section 4, Table 2). All measurements
 651 use the script’s default eval batch ($n=2048$ CIFAR-10 test samples) and the training-time B_l
 652 matrices reconstructed from the original training RNG. Layer indices follow the convention
 653 used elsewhere in the paper: $l=0$ is the first residual block (which sees the embedding out-
 654 put) and $l=1..4$ are the deeper residual blocks. The full per-seed values are dumped to
 655 `results/vanilla_dfa_early_ckpts/per_layer_cos_3seed.json`.

Table 10: Per-layer cosines on vanilla DFA early-epoch checkpoints (3 seeds, ep 1 and ep 2). Layer 0 is consistently $\approx +0.42$ across all six measurements while every deep layer (1–4) lies in $[-0.06, +0.02]$, so the headline aggregate Γ on these checkpoints is driven almost entirely by layer 0 even though the deep blocks carry essentially no alignment with the BP gradient.

| seed | ep | $l=0$ | $l=1$ | $l=2$ | $l=3$ | $l=4$ | $\ g_2\ $ |
|------|----|--------|--------|--------|--------|--------|----------------------|
| 42 | 1 | +0.421 | +0.005 | -0.028 | -0.039 | -0.038 | 6.8×10^{-7} |
| 42 | 2 | +0.437 | -0.002 | -0.040 | -0.055 | -0.054 | 1.6×10^{-7} |
| 123 | 1 | +0.436 | +0.008 | -0.033 | +0.016 | +0.017 | 6.6×10^{-7} |
| 123 | 2 | +0.460 | +0.005 | -0.037 | +0.003 | +0.003 | 1.4×10^{-7} |
| 456 | 1 | +0.418 | +0.011 | -0.026 | +0.007 | +0.006 | 3.8×10^{-7} |
| 456 | 2 | +0.409 | +0.003 | -0.039 | +0.001 | +0.000 | 8.5×10^{-8} |

656 The deep-layer mean across the three seeds at epoch 1 is -0.008 ± 0.013 (matching Table 2), and
 657 at epoch 2 is -0.018 ± 0.018 . Layer 0 stays at $+0.42 \pm 0.02$ across all six measurements, so the
 658 layer-0-dominance pattern is not a single-seed coincidence: it is consistent across seeds and across
 659 the early epochs in which $\|g_2\|$ remains above the 10^{-7} diagnostic-(b) floor. This is the per-seed
 660 evidence behind the Section 4 claim that aggregate cosine on vanilla DFA can look mildly positive
 661 only because layer 0 carries the entire alignment budget.

662 L Reproducibility

663 All headline audit results in the main text should be reported over the locked seed set $\{42, 123, 456\}$,
 664 with the same seed bundle reused across methods wherever possible so that between-method compar-
 665 isons are not driven by different data orders or initialization luck. Every released result table
 666 should specify the architecture, optimizer, learning-rate schedule, batch size, augmentation recipe,
 667 number of epochs, checkpoint selection rule, and whether each diagnostic was measured at the final
 668 checkpoint or along a stored temporal trajectory.

669 Hyperparameters should be listed exactly as run, not reconstructed from memory after the fact. For
 670 intervention experiments, the appendix should report the penalty coefficient, where in the network
 671 the penalty is applied, and which control runs share the same added objective. For diagnostic scripts,
 672 reproducibility requires logging the model mode, minibatch identity, and layer-index convention
 673 used for per-layer statistics. The point of this appendix is simple: because the paper’s claims hinge
 674 on how evaluation is performed, measurement configuration is part of the result and must be repro-
 675 ducible with the same care as training configuration.