GRADIENT TOXICITY: WHEN AUXILIARY HEADS COLLAPSE THE SHARED ENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-task optimization evaluates how tasks share gradients, but task compatibility does not determine whether the shared representation survives. We introduce *gradient toxicity* as a second diagnostic axis: the signed contribution of an auxiliary update to representation-variance contraction. An exact variance identity shows why conflict information alone cannot certify survival, and ten multi-task optimizers collapse in $10/10$ seeds at the tested high doses while detachment collapses in none. Because the axis is measured by a cheap frozen probe, the validity of that probe must be established rather than assumed. We derive a computable margin condition under which the probe agrees with the instantaneous variance derivative, and decompose its failure into restoration misalignment, encoder geometry, and mismatch between the probe and the operating gradient. Under one controlled toxifier, predicted doses track measured scatter crossings on battery data but overestimate them by 30–$116\times$ on RadioML and by 3.9–$32.6\times$ on NYUv2, and changing only the calibration time switches a HAR filter from $10/10$ to $0/10$ collapses. Real auxiliary heads separate unsigned radial loading from signed contraction, and representation survival from task utility: on NYUv2, depth RMSE improves from 1.33 to 1.09 as the representation contracts. Under a corrected cell-disjoint battery protocol, detachment reduces state-of-health RMSE from 5.57 to 2.10 percentage points. Selective filtering retains a useful auxiliary effect where detachment removes it, but not at every dose. These results establish representation survival as a separate evaluation requirement and probe validity as a quantity to audit.

1 INTRODUCTION

A shared representation is the central resource of multi-task learning. Yet the usual question—whether task gradients cooperate or conflict—does not ask whether that resource remains informative. A task head can compensate for shrinking features, and a low task loss can coexist with a nearly constant encoder. On our controlled HAR benchmark, accuracy remains near 0.95 at only 0.05 of the reference scatter. On NYUv2, depth RMSE improves from 1.33 to 1.09 while relative scatter falls to 0.04. Task performance alone therefore leaves a gap in the evaluation of representation learning.

Conflict resolution addresses a different question. PCGrad (Yu et al., 2020), CAGrad (Liu et al., 2021a), and related methods combine task gradients using directional or loss-based criteria. Representation variance, however, changes according to the update's projection onto the *variance gradient*. We call the auxiliary contribution to this change *gradient toxicity*. The distinction is structural: two auxiliary gradients can have identical task-gradient inner products and norms, yet drive variance in opposite directions. In a contraction-dominated regime, satisfying a conflict criterion need not preserve the representation. Ten widely used multi-task optimizers exhibit this gap: each collapses in all ten seeds at the tested high doses on three domains, while detachment collapses in none (Figure 1b).

Adding a second axis also raises a question about the diagnostic itself: *when should it be trusted?* A projected-gradient norm is attractive because it is cheap to measure before auxiliary training. But its interpretation depends on what its numerator measures, whether its denominator represents restoration, and whether the encoder and auxiliary remain close to the calibration state. A convincing diagnosis must expose these dependencies, including the circumstances in which the probe fails.

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.

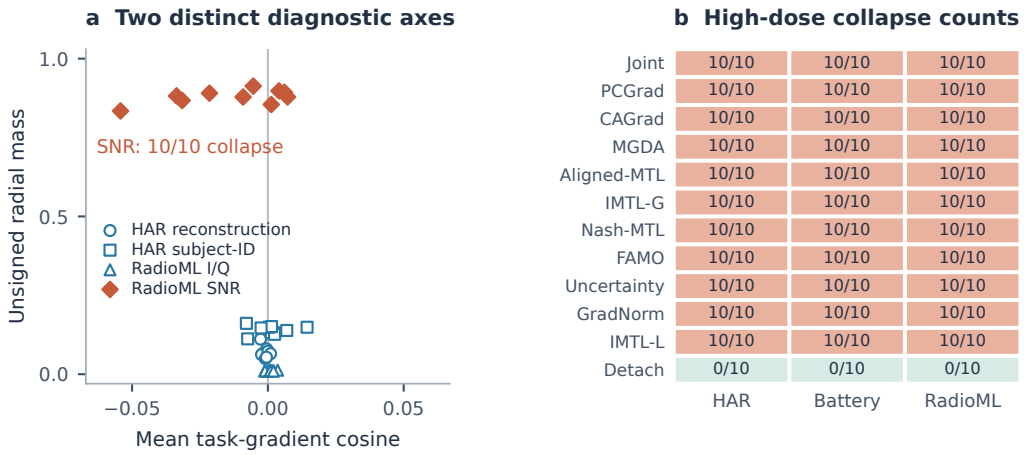

Figure 1: (a) Terminal probes of four real auxiliary heads, ten seeds each. The horizontal coordinate is the *mean per-example* task-gradient cosine; the vertical coordinate is unsigned radial loading. SNR regression has high loading and $10/10$ collapsed endpoints. (b) Operational collapse counts at controlled doses 0.1 (HAR), 3000 (battery), and 10000 (RadioML). All ten optimizers retain collapse. The detached reference uses separate ten-seed controls. Full metrics and control settings are in Table 1.

Starting from an exact chain-rule identity, we derive a margin condition that separates trustworthy local probe decisions from unresolved ones. The condition decomposes the discrepancy into restoration direction, encoder-induced geometry, and probe mismatch. It also predicts distinct *types* of error: a large primary-gradient norm can conceal weak restoration, an unsigned projection loses the sign of contraction, and a calibration statistic can drift even without auxiliary contact.

Our experiments separate mechanism controls from real auxiliary objectives. With the same fixed variance toxifier, the frozen probe estimates a scatter crossing within a factor of 3.1 on battery data but overestimates it by 30–$116\times$ on RadioML. Thus even a designed contracting loss is insufficient to validate a probe. Calibration-time interventions on HAR reproduce false negatives, while zero-dose controls expose large probe drift without auxiliary-induced damage. Real heads add sign ambiguity and demonstrate that passing a scatter threshold does not imply recovery of predictive utility.

The contributions are threefold. First, we formalize representation survival as an evaluation axis distinct from task-gradient conflict. Second, we characterize a computable local validity boundary for a frozen toxicity probe and test observable failure modes across domains and calibration conditions. Third, we use detachment and spectral toxicity filtering (STF) as interventions that test the diagnosis, distinguishing survival, auxiliary utility, and task recovery. The resulting benchmark reports both the probe and the conditions under which its interpretation holds.

2 A Second Axis: Representation Survival

Let $H(\theta) \in \mathbb{R}^{B\times m}$ contain the outputs of a shared encoder on a fixed diagnostic batch. Training minimizes $L_p + \alpha L_a$, with primary loss L_p, auxiliary loss L_a, and dose $\alpha \geq 0$; multiple auxiliaries can be absorbed into their weighted sum. Write $G_p = \nabla_\theta L_p$ and $G_a = \nabla_\theta L_a$. Define

$$V(\theta) = \frac{1}{2B}\|C_B H(\theta)\|_F^2, \qquad C_B = I - \mathbf{1}\mathbf{1}^\top/B, \qquad v = \nabla_\theta V. \tag{1}$$

Here $2V$ is the covariance trace with divisor B: a rotation-invariant measure of variation across examples. Under gradient flow,

$$\dot{V} = r - \alpha c, \qquad r = -v^\top G_p, \qquad c = v^\top G_a. \tag{2}$$

We call $c > 0$ *auxiliary contraction*, and $r > 0$ *primary restoration*. When both are positive, $\alpha c/r > 1$ defines the instantaneous contraction-dominated regime. Reporting (r, c) also covers a primary update that itself reduces variance. This identity retains the encoder Jacobian; it defines a signed diagnostic without a reduced dynamical model.

Proposition 1 (Conflict information does not identify contraction)**.** *The task-gradient Gram matrix does not determine the sign of $\dot{V}$. In particular, there are two auxiliary updates with identical task-gradient norms and conflict, but opposite effects on variance under the same primary update.*

Proof. Take $G_p = e_1$, $G_a^{\pm} = \pm e_2$, and $v = -e_1 + \beta e_2$ for $\beta > 0$. Both pairs have identity Gram matrix and zero conflict, whereas $r = 1$ and $c^{\pm} = \pm\beta$. For $\alpha\beta > 1$, $\dot{V}^+ < 0 < \dot{V}^-$. These gradients are locally realizable by a smooth encoder and losses (Appendix A). □

The proposition identifies the missing information: a rule based only on task-gradient relations has no variance constraint. Reducing the effective auxiliary dose can preserve variance, but conflict resolution alone provides no survival certificate for a retained auxiliary update. The high-dose comparison tests this gap at fixed operating points.

Equation (2) describes instantaneous total variation. Eventual collapse also depends on how this balance evolves. In the symmetric linear model $\dot{z} = (R - \alpha A)z$, $R \succ 0$, $A \succeq 0$, some mode contracts when $\alpha\lambda_{\max}(R^{-1/2}AR^{-1/2}) > 1$, whereas *all* modes contract only when the corresponding minimum eigenvalue exceeds $1/\alpha$ (Theorem 3). A task head can exploit surviving modes even as other directions disappear. We therefore report representation variation and task error together, supplemented by effective rank and prediction variation where available.

3 A Validity Boundary for Toxicity Probes

3.1 From a cheap proxy to a signed diagnostic

Let h and u stack the rows of H and $C_B H$, respectively, and let $g_p = \nabla_h L_p$, $g_a = \nabla_h L_a$. A cheap feature-space probe uses the centered radial direction:

Definition 1 (Unsigned radial toxicity proxy)**.** *For $u \neq 0$ and $\|g_p\| > 0$, define*

$$\rho_b = \frac{\|\Pi_u g_a\|}{\|g_p\|}, \qquad \Pi_u g = \frac{u^\top g}{\|u\|^2} u. \tag{3}$$

The experimental estimator $\hat{\rho}$ instead averages regularized per-example ratios. The frozen decision uses a recorded value $q = \hat{\rho}_{t_0}$ and the nominal dose $\alpha_{\text{probe}} = 1/q$.

Two distinctions matter. First, ρ_b is unsigned: contracting and expanding auxiliary directions can have the same value. Second, $\|g_p\|$ measures all primary-gradient directions, including those that do not restore centered variation. Moreover, in the controlled driver the probe differentiates the *isotropic surrogate* $\|C_B H\|_F^2/(Bm)$, while training uses fixed projected-variance auxiliaries. The real-head driver differentiates the actual auxiliary loss. We preserve these estimator choices and audit their consequences; they are not interchangeable measurements of c/r.

3.2 A computable validity margin

The discrepancy can be exposed without assuming radial training dynamics. Let $J = \partial h/\partial\theta$, $K = JJ^\top$, $e = u/\|u\|$, and $n_p = \|g_p\| > 0$. At an audited checkpoint choose any $k > 0$ and write $K = k(I + E)$. Define

$$a = \frac{e^\top g_a}{n_p}, \quad d = -\frac{e^\top g_p}{n_p}, \quad \epsilon_a = \frac{e^\top E g_a}{n_p}, \quad \epsilon_p = -\frac{e^\top E g_p}{n_p}. \tag{4}$$

Here a is signed radial loading relative to the primary norm, d measures restoration alignment, and the ϵ terms measure the effect of encoder geometry. For the actual auxiliary, $|a| = \rho_b$. Products with K can be evaluated by Jacobian–vector and vector–Jacobian products, without constructing the full matrix.

Theorem 1 (Local validity boundary of a frozen probe)**.** *For any frozen probe value $q \geq 0$, the variance dynamics satisfy*

$$\frac{\dot{V}}{N} = 1 - \alpha q + \mathcal{E}, \qquad N = \frac{k\|u\|n_p}{B} > 0, \qquad \mathcal{E} = (d-1) + \epsilon_p - \alpha(a - q + \epsilon_a). \tag{5}$$

Consequently, with

$$\Delta = |d-1| + |\epsilon_p| + \alpha\big(|a-q| + |\epsilon_a|\big), \qquad |1 - \alpha q| > \Delta, \tag{6}$$

the probe and the exact variance derivative agree in sign. If $1 - \alpha q > \Delta$, variance increases; if $1 - \alpha q < -\Delta$, it decreases. The remaining region requires the signed audit.

Proof. The chain rule gives $\dot{V} = -u^\top K(g_p + \alpha g_a)/B$. Substitute $K = k(I + E)$, divide by N, and add and subtract $1 - \alpha q$. The triangle inequality bounds $|\mathcal{E}|$ by Δ. □

The theorem supplies a *sufficient, checkpoint-local* boundary. Its decomposition attributes a scalar threshold's discrepancy to quantities that can be checked separately. The directly evaluated $\mathcal{E}$ is sharper than the conservative envelope Δ. A uniform bound over a trajectory would transport the decision through time; the archived neural-network experiments do not measure such a bound. The checkpoint-local certificate itself is audited on 405 HAR checkpoints in Appendix A.

In the identity metric ($E = 0$) with positive signed loading, $\dot{V}/N = d - \alpha a$. If $a = q$ but $0 < d < \alpha q < 1$, the frozen gate reports a subcritical dose although variance contracts: the primary norm overstates restoration. If $a = -q$, the unsigned gate can report danger while the auxiliary expands variance. If a changes after calibration, the error moves with $a - q$. These are algebraic predictions about the sign of the diagnostic error, rather than an assumption that every high probe value indicates collapse.

The mismatch term also separates into actionable components. If a_0 is signed loading at calibration for the actual auxiliary, then

$$|a_t - q| \leq |a_t - a_0| + |a_0 - |a_0|| + \big||a_0| - q\big|. \tag{7}$$

The three terms isolate temporal drift, loss of sign, and the surrogate/aggregation discrepancy. This identifies the scope of the familiar radial approximation: a contracting auxiliary, a predominantly restoring primary gradient, a compatible encoder metric, and a transferable calibration statistic. A fixed quadratic toxifier simplifies only part of this list.

A scalar reduction of the same balance, $\dot{\sigma} = -\alpha\kappa\sigma + b(\sigma)$, has a saddle-node threshold $\alpha_c = \lambda_\star/\kappa$ when $b(\sigma)/\sigma$ has a unique nondegenerate maximum $\lambda_\star$ and the stated endpoint behavior (Theorem 2). Identifying $1/q$ with that threshold additionally requires a calibration relation to $\kappa/\lambda_\star$. We test this transfer using empirical half-height crossings, denoted $\alpha_{1/2}$, separately from the operational collapse threshold. This distinction turns disagreement into evidence about the probe, rather than relabeling a fitted crossing as a proved bifurcation.

4 An Empirical Map of the Boundary

4.1 Protocols and observables

The controlled benchmark spans UCI-HAR (Anguita et al., 2012), RadioML2016.10A (O'Shea et al., 2016), battery health (Saha & Goebel, 2007), and NYUv2. It uses a three-layer MLP of width m and a fixed auxiliary $L_{\text{tox}} = K_a^{-1}\sum_{k=1}^{K_a} \mathrm{Var}_b(W_k h)$, with $K_a = 3$ seeded projections. Adam uses learning rate 10^{-3} and batch size 256; experiments use 5–10 seeds and bf16 autocast. The default frozen probe follows four detached warm-up epochs. Detailed data partitions, encoder variants, and real-head protocols are in Appendices F–H.

We retain the archived operational criterion: collapse means relative scatter $s = \mathcal{S}_h/\mathcal{S}_{\text{single}} < 0.3$, with the same-seed reference and

$$\mathcal{S}_h = \mathbb{E}_b\Big[\mathrm{std}_j\Big(h_j^{(b)} - \mathbb{E}_{b'} h_j^{(b')}\Big)\Big]. \tag{8}$$

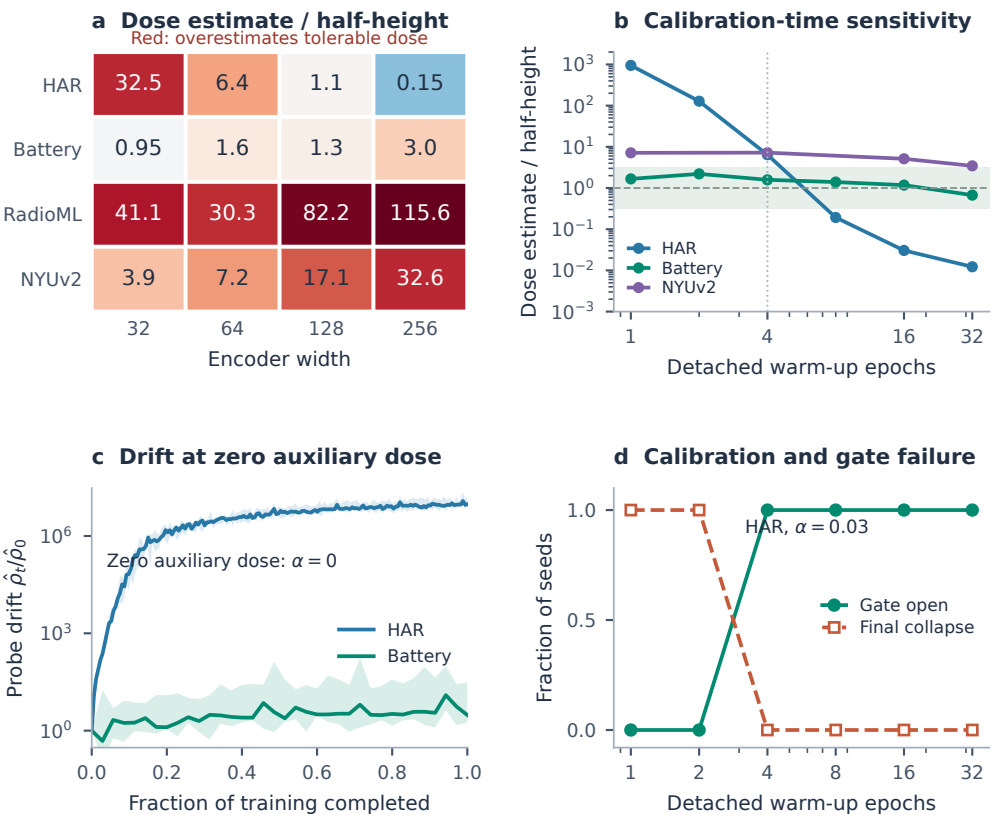

Figure 2: (a) $\alpha_{\text{probe}}/\alpha_{1/2}$ for 16 domain–width cells; *all use the controlled toxifier*. Red indicates an optimistic dose estimate, blue a conservative one. (b) Changing calibration time changes this ratio at $m = 64$; the shaded factor-3 band is a visual reference, not a confidence interval. Both panels use ten-seed mean probes. (c) At $\alpha = 0$, HAR's probe drifts by orders of magnitude while battery's remains comparatively stable; median and full range over five seeds, each normalized to its initial value. (d) HAR at $\alpha = 0.03$: early calibration leaves all gates closed and all ten runs collapse; at $W \geq 4$, all gates open and none collapse. Source values and crossing definitions are in Table 10 and Appendices H.2–H.3.

This coordinate-dependent proxy measures a different observable from V and can miss variation in the all-ones feature direction. Accordingly, "collapse" in the empirical counts means this specified threshold. The half-height $\alpha_{1/2}$ interpolates the *mean* dose–scatter curve at $s = 0.5$ in log dose. It is a reproducible transition summary, not an estimate of complete information loss. Error bars or bands denote population standard deviation across the stated seeds unless specified otherwise.

4.2 A CONTROLLED TOXIFIER DOES NOT GUARANTEE A VALID PROBE

Figure 2a replaces a pooled prediction score with the domain–width error map. Battery ratios lie between 0.95 and 3.01. HAR changes from 32.5 at $m = 32$ to 1.11 at $m = 128$ and 0.15 at $m = 256$, reversing the direction of bias. RadioML overestimates the crossing in every width by 30–116×; NYUv2 does so by approximately 3.9–32.6×. A single median hides both systematic optimism and a reversal with width.

These results locate an empirical transfer boundary *within* the controlled family. They do not partition synthetic and real objectives: all sixteen cells use fixed projected-variance auxiliaries, and their probes use the isotropic surrogate. The relevant distinction is whether the probe captures the operating contraction–restoration balance. The near-match on battery and wide HAR is useful calibration evidence; the RadioML bias shows that the same estimator and training recipe do not transfer uni-

formly. The algebra in Section 3 identifies candidate discrepancies to measure, while the archived cells establish the errors rather than their unique causal allocation.

4.3 Calibration time predicts a reproducible gate failure

Holding the data, encoder, and probe definition fixed, we vary the detached warm-up length W. On HAR at $m = 64$, $1/\bar{\hat{\rho}}$ moves from 1.34 at $W = 1$ to 0.181 at $W = 2$, 0.00907 at $W = 4$, and 1.7×10^{-5} at $W = 32$ (Figure 2b). The measured half-height is 0.00141. The default $W = 4$ prediction lies in the coarse dose bracket $[0.001, 0.01]$, but differs from the interpolated crossing by $6.4\times$; the two statements answer different questions.

The calibration change has a direct intervention consequence. At dose $\alpha = 0.03$, $W = 1$ and 2 leave all ten gates closed; all runs end below the scatter criterion. For $W \geq 4$, all gates open and none collapse, with relative scatter 0.99–1.00 (Figure 2d). At $\alpha = 100$, all six warm-up choices open the gate and avoid collapse. Thus a successful high-dose repair can conceal a calibration failure near the operating boundary. The gate arithmetic predicts activation before auxiliary contact; the subsequent outcomes test whether that activation is sufficient in this protocol.

4.4 Continuous re-probing can create the opposite error

Recomputing the same norm ratio is not a universal correction. On unfiltered HAR trajectories, the probe increases by orders of magnitude, including at $\alpha = 0$, when no auxiliary update reaches the encoder (Figure 2c). At positive doses, a running gate activates in all seeds of ten HAR configurations whose *final* relative scatter is 0.50–1.04. These are false alarms relative to the endpoint criterion. Battery probes remain $O(1)$; among 35 seeds ending above half-height, only one gate activation occurs, at the last step (Appendix H.2).

The zero-dose control distinguishes drift of the statistic from damage caused by the auxiliary. Because the primary-gradient norm is in the denominator, improved task fitting can inflate the ratio without establishing signed contraction. We evaluate every term of Theorem 1 on 405 HAR checkpoints; the median restoration alignment is 0.065 rather than 1. The appropriate response is to inspect the signed flux and its components, rather than to interpret each increase of $\hat{\rho}$ as a loss of representation survival.

4.5 Real heads expose sign and intervention mismatch

Terminal probes of four real heads provide a second stress test (Figure 1a). Across ten seeds, mean per-example task-gradient cosines have magnitude at most 0.06, while SNR regression places 88% of its gradient norm on each example's radial axis, compared with 1–14% for the other heads. This separates the collapsed SNR endpoints from benign endpoints. It is not evidence that every example has negligible conflict: the SNR head has negative primary–auxiliary dot products on approximately 54% of probe examples.

The signed measurement is equally informative. Despite its large unsigned loading, the SNR head's mean signed radial cosine is approximately -0.12 at the endpoint. An unsigned projection therefore cannot be read as a signed contraction rate or a forecast from that state. Both observations support the second axis while motivating its validity audit. The result also explains why success on a deliberately contracting objective is insufficient evidence of transfer to an arbitrary real auxiliary loss.

5 Spectral Toxicity Filtering as a Diagnostic Test

Detachment blocks all auxiliary gradients to the encoder and is the reference intervention when the auxiliary adds no primary-task value. STF instead restricts the auxiliary channel. Its idealized backward form removes a radial projection,

$$F_w[g_a] = g_a - w\Pi_u g_a, \qquad w \in [0, 1], \tag{9}$$

while the implemented selective arms transform the auxiliary input, $\widetilde{H} = H - wC_B HP$, with a frozen projector P or a centered radial direction. This changes both the forward input and its

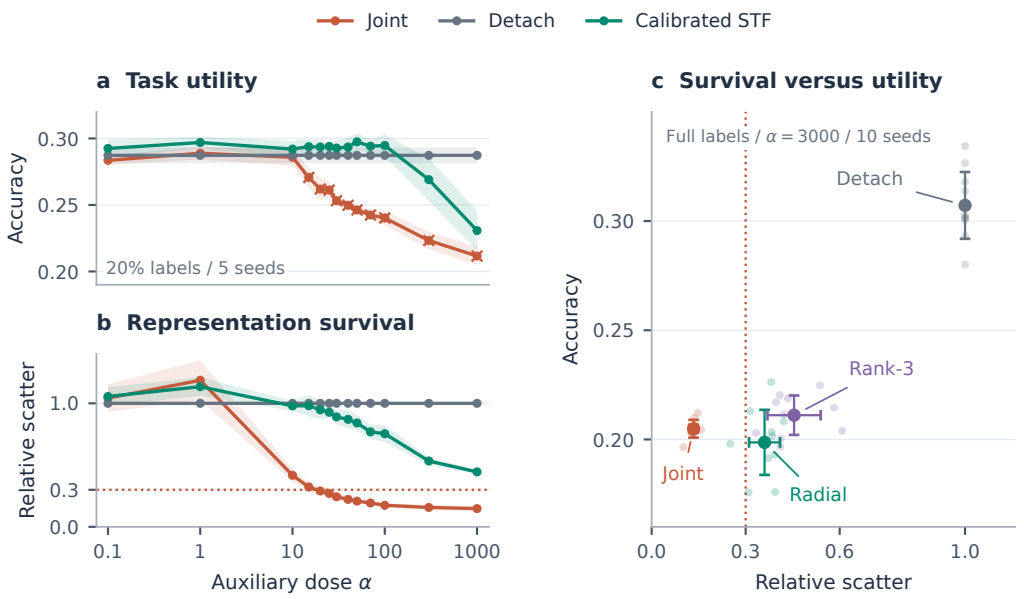

Figure 3: Real RadioML SNR-regression auxiliary. (a,b) At 20% primary labels and width 64, calibrated STF retains a mean accuracy advantage over detachment at intermediate doses; the advantage disappears at extreme doses even while scatter remains above 0.3. Bands are seed standard deviations; crosses mark doses with at least one collapsed joint run. Five unique seeds per dose; duplicate log records are removed. (c) A separate fully labeled, dose-3000 protocol gives a different ordering: rank-3 filtering clears the scatter criterion in all ten seeds, but detachment has higher accuracy. Faint points are seeds; large points and bars show means and standard deviations. The dotted line is the 0.3 scatter criterion.

gradient; the auxiliary head itself continues training. Appendix E specifies the forward and backward semantics.

The calibrated arm freezes its warm-up probe and opens the gate when $\alpha\hat{\rho} \geq 1$. The rank variant selects

$$r_{\text{sel}} = \min\{r \leq R : \alpha\hat{\rho}_{\text{res}}[r] < 1\}, \tag{10}$$

using nested candidate subspaces; if none passes, it uses the cap and records the unresolved residual. The radial arm is a separate intervention. Theorem 5 characterizes minimum deletion rank for a specified PSD operator; empirical PCA bases require their own identification. This makes gate calibration and subspace choice two distinct parts of the intervention.

A useful auxiliary can also become toxic: in the 20%-label SNR dose scan (Figure 3a,b), joint training has a healthy encoder at low dose, but collapses in every seed from $\alpha = 30$ onward. Calibrated STF avoids collapse at every sampled dose. Over $\alpha \in [15, 100]$, its mean accuracy is approximately 0.5–1.0 percentage points above detachment. These five-seed results establish a tested regime in which selective intervention preserves a useful effect; they do not imply a uniform advantage across protocols.

The fully labeled SNR experiment demonstrates why the benchmark needs both axes of evaluation (Figure 3c). Detachment reaches 0.307 accuracy with $0/10$ collapses. The radial arm reaches 0.199 with $1/10$ collapses, while the rank-3 arm reaches 0.211 with $0/10$. Removing enough contraction to clear a scatter threshold does not ensure that the remaining auxiliary update helps the task. In a separate HAR reconstruction study, an auxiliary-gradient basis uses 30.75 directions versus PCA's 62.0 among the same four gate-open seeds at similar accuracy (Table 4); basis choice changes how much of the auxiliary channel is removed.

At the tested high doses, calibrated filtering changes collapse counts from $10/10$ to $0/10$ across four controlled scientific domains. Protection also transfers to a Transformer encoder: at RadioML

Algorithm 1 Auditing representation survival and probe validity

Require: Diagnostic batch, healthy reference, checkpoint θ, operating losses and dose α, recorded calibration value q.

1: Report task metrics, covariance trace $2V$, effective rank, and the declared operational scatter criterion.
2: Differentiate V and the actual losses; report signed restoration r and auxiliary contraction c from (2).
3: Record the cheap decision $1 - \alpha q$ and compute the discrepancy $\mathcal{E}$ and envelope Δ using (5)–(6).
4: **if** $|1 - \alpha q| > \Delta$ **then**
5: Mark the probe decision as locally certified for this batch and update model.
6: **else**
7: Use the signed audit; report restoration misalignment, geometry, and probe mismatch separately.
8: **end if**
9: Compare joint training, detachment, and any selective intervention using both survival and task utility at matched doses and seeds.

$\alpha = 3000$, joint training collapses in $10/10$ seeds at accuracy 0.156, whereas calibrated filtering has $0/10$ collapses and relative scatter 0.970. On CelebA's Smiling task (Liu et al., 2015), the calibration probe spans $\hat{\rho} = 1602$–3585 across widths; at $\alpha = 100$, joint training collapses in $5/5$ seeds and calibrated filtering in $0/5$. This intervention result coexists with the probe errors in Figure 2: an extreme dose can open a biased gate reliably. Filtering before a conflict operator also preserves scatter in the tested compositions (Table 21). Variance regularization offers another intervention, but its ability to clear the scatter threshold and its task accuracy vary independently (Appendix I).

On the corrected, cell-disjoint MATR battery protocol (Severson et al., 2019), validation-selected detachment improves SOH RMSE from 5.57 to 2.10 percentage points and observed-RUL MAE from 81.5 to 46.7 cycles. Selective filtering does not improve on detachment there (Table 2). The historical ranking reverses after changing evaluation-selected checkpoints and censoring-blind scoring together, so the corrected comparison concerns the complete protocol. Financial and battery application details remain in the appendix; their role is to test diagnostic scope, not to redefine survival through whichever task metric improves.

6 A Diagnostic Standard for Shared Representations

The practical output of the analysis is a reporting protocol, not a universal cutoff on $\hat{\rho}$. It separates three questions: whether an update contracts the chosen observable, whether a cheap probe captures that update, and whether the resulting representation supports the task. Each question has its own measurable answer (Algorithm 1).

The certificate is sufficient, so a large Δ does not establish that a probe is wrong. Directly evaluating the signed discrepancy resolves the current checkpoint; attributing it to the terms in (4) determines the next measurement. Weak restoration is addressed by a directional primary reference, a sign discrepancy by a signed numerator, a surrogate discrepancy by the actual auxiliary gradient, and temporal mismatch by a new calibration audit. Merely recomputing the same unsigned norm ratio does not address all four cases, as the zero-dose control demonstrates.

A survival intervention should first alter the contraction observable in the intended direction, then preserve or improve predictive utility. These criteria can disagree: the rank-3 SNR filter passes the scatter criterion while underperforming detachment, and the low-label radial filter loses its accuracy advantage at extreme doses. Reporting only collapse counts would hide both outcomes. Likewise, a gate that opens correctly can still choose an ineffective deletion subspace. The diagnostic benchmark therefore retains closed-gate failures, open-gate failures, and successful protection as different outcomes.

This protocol can be added to an existing multi-task optimizer without replacing its objective. The cheap probe provides a screening measurement; the signed audit checks its interpretation at selected

operating points. For adaptive optimizers, the actual proposed parameter step replaces the gradient-flow direction, with the finite-step remainder in Appendix A. An early-warning claim additionally requires calibration before the outcome and evaluation at declared future horizons. Our endpoint associations, warm-up interventions, and local certificate are reported separately so each supports the question it actually measures.

7 Related Work

Multi-task learning (Caruana, 1997) studies transfer among objectives, and task-grouping studies ask when joint training helps (Standley et al., 2020; Vandenhende et al., 2022). PCGrad and CA-Grad address gradient conflict (Yu et al., 2020; Liu et al., 2021a); MGDA, Nash-MTL, and FAMO pursue Pareto, bargaining, or progress-balancing criteria (Sener & Koltun, 2018; Navon et al., 2022; Liu et al., 2023). Other methods modify task-gradient geometry or loss weights (Senushkin et al., 2023; Liu et al., 2021b; Chen et al., 2018; Kendall et al., 2018; Wang et al., 2021; Limarenko et al., 2026; Zhou et al., 2025). Our structural result concerns what task-gradient relations leave unidentified: their effect on representation variance. It motivates an additional diagnostic alongside these optimization criteria.

Representation collapse has been studied extensively in self-supervised learning, including stop-gradient mechanisms (Grill et al., 2020; Chen & He, 2021; Caron et al., 2021), variance and covariance regularization (Zbontar et al., 2021; Bardes et al., 2022; Ermolov et al., 2021), and spectral dynamics (Jing et al., 2022; Dong et al., 2021; Kim et al., 2025). RankMe evaluates learned representations through effective rank (Garrido et al., 2023). Here collapse refers to reduced variation across examples, distinct from terminal classifier geometry in neural collapse (Papyan et al., 2020; Wang et al., 2026). Our contribution connects an update-level signed observable to the validity of the cheap probe used to approximate it.

Auxiliary-gradient projection is also established. ATTITTUD and GradOPS choose subspaces from task influence or task-gradient spans (Dery et al., 2021; Zhu et al., 2025); other approaches select or combine auxiliary tasks (Navon et al., 2021; Jiang et al., 2023; Shin et al., 2024). Our time-series domains are activity recognition (Anguita et al., 2012), modulation recognition (O'Shea et al., 2016), and battery health (Saha & Goebel, 2007; Severson et al., 2019; Attia et al., 2020; Pradhan & Chakraborty, 2022), where auxiliary learning is common (Zhang & Yang, 2022). We use filtering to test the contraction diagnosis and compare it with detachment. The contribution is the diagnosis and its scope: what must be measured to interpret a toxicity probe, why a threshold can transfer in one regime and fail in another, and how to evaluate the intervention without conflating survival with task utility.

8 Conclusion

Task-gradient compatibility does not certify that a shared representation survives. Gradient toxicity adds the missing variance-sensitive axis, and its exact signed form exposes when a cheap frozen probe is reliable. We give a computable local validity boundary and identify distinct failure patterns under calibration changes, training-state drift, and real-head geometry. The diagnostic's scope is determined by these mechanisms. Detachment and selective filtering test the consequences: a representation can survive without recovering task utility, and an auxiliary can be useful before becoming toxic. Multi-task evaluation should report both outcomes and audit the probe that connects them.

AI USE STATEMENT

Generative AI tools assisted with drafting, language editing, code, literature retrieval, proof development, and figure design. The new diagnostic identities were checked with executable numerical examples; assumptions are stated explicitly with each claim and in Appendix D. Figures are drawn from recorded experimental results. The authors remain responsible for verifying the final text, claims, code, and artifacts.

ETHICS STATEMENT

We use existing datasets: UCI-HAR, RadioML2016.10A, the NASA PCoE and MATR battery archives, public A-share market data, NYUv2, and CelebA. We collected no new human-subject data. The HAR identity auxiliary is evaluated as a training-head probe, not a held-out privacy attack. Battery prediction is safety-relevant; Section 4.4 describes the observed limitations, and deployment requires task-specific validation.

REPRODUCIBILITY STATEMENT

Appendices D–E give the assumptions and scope of the theoretical results. The repository provides experiment code, per-seed records, counterexample checks, generator audits, and the corrected battery validation protocol. Historical battery results selected checkpoints on the evaluation split and are exploratory, rather than independent test estimates. Code and records are available in the anonymized repository: `https://github.com/code-2026l/RepFix`.

REFERENCES

Ahmed Allibhoy and Jorge Cortés. Control barrier function based design of gradient flows for constrained nonlinear programming. *IEEE Transactions on Automatic Control*, 69(6):3499–3514, 2024. doi: 10.1109/TAC.2023.3306492.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International Workshop on Ambient Assisted Living (IWAAL)*, volume 7657 of *Lecture Notes in Computer Science*, pp. 216–223. Springer, 2012.

Peter M. Attia, Aditya Grover, Norman Jin, Kristen A. Severson, Taylor M. Markov, Yu-Hsin Liao, Michael H. Chen, Bryan Cheong, Nicholas Perkins, Zi Yang, Patrick K. Herring, Muratahan Aykol, Stephen J. Harris, Richard D. Braatz, Stefano Ermon, and William C. Chueh. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578(7795): 397–402, 2020.

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. SCARF: Self-supervised contrastive learning using random feature corruption. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning (ICML)*, 2018.

Lucio M. Dery, Yann Dauphin, and David Grangier. Auxiliary task update decomposition: The good, the bad and the neutral. In *International Conference on Learning Representations (ICLR)*, 2021.

Shibhansh Dohare, Juan Hernandez-Garcia, Qingfeng Lan, Parash Rahman, Ashique Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. TSMixer: Lightweight MLP-Mixer model for multivariate time series forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning (ICML)*, 2021.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Junguang Jiang, Baixu Chen, Junwei Pan, Ximei Wang, Dapeng Liu, Jie Jiang, and Mingsheng Long. ForkMerge: Mitigating negative transfer in auxiliary-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Donghyeon Kim, Chae-Bong Sohn, Do-Yup Kim, and Dae-Yeol Kim. A taxonomy and theoretical analysis of collapse phenomena in unsupervised representation learning. *Mathematics*, 13(18): 2986, 2025.

Lubor Ladický, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 89–96, 2014.

Evgeny Alves Limarenko, Anastasiia Studenikina, Svetlana Illarionova, and Maxim Sharaev. GCond: Gradient conflict resolution via accumulation-based stabilization for large-scale multi-task learning. *IEEE Access*, 14:42086–42104, 2026.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 57226–57243, 2023.

Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2021b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Aviv Navon, Idan Achituve, Haggai Maron, Gal Chechik, and Ethan Fetaya. Auxiliary learning by implicit differentiation. In *International Conference on Learning Representations (ICLR)*, 2021.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning (ICML)*, 2022.

Timothy J. O'Shea, Johnathan Corgan, and T. Charles Clancy. Convolutional radio modulation recognition networks. In *Engineering Applications of Neural Networks (EANN)*, volume 629 of *Communications in Computer and Information Science*, pp. 213–226. Springer, 2016.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117 (40):24652–24663, 2020.

Sunil K. Pradhan and Basab Chakraborty. Battery management strategies: An essential review for battery state of health monitoring techniques. *Journal of Energy Storage*, 51:104427, 2022.

Bhaskar Saha and Kai Goebel. Battery data set. NASA Ames Prognostics Data Repository, 2007. URL `https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/`.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Kristen A. Severson, Peter M. Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H. Chen, Muratahan Aykol, Patrick K. Herring, Dimitrios Fraggedakis, Martin Z. Bazant, Stephen J. Harris, William C. Chueh, and Richard D. Braatz. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.

Seokwon Shin, Hyungrok Do, and Youngdoo Son. Learning representation for multitask learning through self-supervised auxiliary learning. In *European Conference on Computer Vision (ECCV)*, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012.

Trevor Standley, Amir Zamir, Dong Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning (ICML)*, 2020.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. Addison-Wesley, Reading, MA, 1994.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1): 24–36, 1958.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3614–3633, 2022.

Youjun Wang, Boqi Li, Xin Zou, and Weiwei Liu. Neural collapse in multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations (ICLR)*, 2021.

Greg Yang and J. Edward Hu. Feature learning in infinite-width neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Injecting imbalance sensitivity for multi-task learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7236–7244, 2025.

Shijie Zhu, Hui Zhao, Tianshu Wu, Pengjie Wang, Hongbo Deng, Jian Xu, and Bo Zheng. Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 204–212, 2025.

A Proofs and Audit Details for the Diagnostic Boundary

Let $\theta \in \mathbb{R}^2$, $v_0 = -e_1 + \beta e_2$, and $z(\theta) = v_0 + \theta$. Choose a two-example encoder $H(\theta) = (z^\top, -z^\top)^\top$. Then $V = \|z\|^2/2$ and $\nabla_\theta V = v_0$ at $\theta = 0$. The losses $L_p = \|z - (v_0 - e_1)\|^2/2$ and $L_a^\pm = \|z - (v_0 \mp e_2)\|^2/2$ are nonnegative smooth functions of the representation. Their gradients at zero are e_1 and $\pm e_2$. Thus Proposition 1 concerns realizable encoder/loss configurations, not arbitrary incompatible gradients. Both task pairs have the same Gram matrix, while the signed variance derivatives differ.

Vectorization of the centered batch gives $u = (C_B \otimes I_m)h$ and $\nabla_\theta V = J^\top u/B$. Hence $\dot{V} = -u^\top J J^\top (g_p + \alpha g_a)/B$. For any fixed positive scale k, write $JJ^\top = k(I + E)$. This is a decomposition, not an assumption of isotropy. With the definitions in (4), the normalized derivative is $d + \epsilon_p - \alpha(a + \epsilon_a)$, which proves (5). Taking absolute values term by term proves (6). The exact discrepancy $\mathcal{E}$ can be smaller than Δ because signed terms cancel. Failure of the sufficient margin condition therefore leaves the probe undecided; it does not prove the probe wrong. Choosing k must be reported, since it changes the decomposition and the tightness of this conservative envelope, although it cannot change $\dot{V}$.

Record the batch, checkpoint, losses, auxiliary dose, probe value q, and scale k. Compute g_p, g_a for the actual operating losses and u from the unfiltered representation. Evaluate Kg by first computing $J^\top g$, then a Jacobian–vector product with J; a dense K is unnecessary. This yields $a, d, \epsilon_a, \epsilon_p$, the exact signed flux, and the validity margin. When $u = 0$ or $\|g_p\| = 0$, the normalized decomposition is undefined and the unnormalized identity (2) remains the appropriate diagnostic. This audit costs more than a frozen feature-space probe. Its purpose is to identify why the cheap decision can or cannot be transported to an operating checkpoint. The archive used here contains the cheap proxies and intervention outcomes. On 405 HAR checkpoints the frozen probe agrees in sign with the exact variance derivative on 51% of them, while the certificate is never wrong when it fires: 2 checkpoints at the zero-trace scale $k = \mathrm{tr}(K)/(Bm)$, and 15 under a per-checkpoint optimal one.

For $K = I$, the local condition is $d - \alpha a < 0$. Take $d = 0.1$, $a = q = 0.2$, and $\alpha = 1$: the nominal gate is closed while the exact derivative is negative. This is a false negative caused by treating the primary norm as radial restoration. Conversely, with $d = 1$, $a = -q = -2$, and $\alpha = 1$, the gate is open but the exact derivative is positive. For a fixed q, the theorem applies at each time with time-indexed components. If $\sup_{t \in I} \Delta_t < |1 - \alpha q|$, its sign conclusion holds throughout the interval I. Obtaining this uniform bound is an additional task; a single frozen observation does not supply it. Even a negative variance derivative on a finite interval does not by itself imply convergence to zero.

Let s_θ be the actual proposed parameter increment, including any optimizer preconditioning, momentum, and weight decay. If ∇V is L-Lipschitz on the segment from θ to $\theta + s_\theta$, then

$$\left|V(\theta + s_\theta) - V(\theta) - v^\top s_\theta\right| \le \tfrac{L}{2}\|s_\theta\|^2. \tag{11}$$

This follows by integrating the gradient along the segment. A first-order margin exceeding this remainder certifies the sign of the finite-step change. The gradient-flow decomposition applies directly to plain gradient updates; the actual-step audit is the corresponding observable for Adam. Separating primary and auxiliary contributions through a nonlinear optimizer requires specifying the counterfactual update and its state. We do not equate a feature-gradient norm ratio with an Adam-step certificate.

The controlled driver computes the probe in `rho_probe` within `cross_domain_mtl.py`:

$$\hat{\rho} = \frac{1}{B}\sum_{i=1}^{B} \frac{\|\Pi_{u_i} \widetilde{g}_{a,i}\|}{\|g_{p,i}\| + 10^{-8}}, \qquad \widetilde{g}_a = \nabla_H \frac{\|C_B H\|_F^2}{Bm}. \tag{12}$$

Training instead uses the model's fixed random projections. Thus a contracting surrogate, per-example averaging, and regularization all enter q. The real-head driver differentiates the specified reconstruction, identity, or SNR loss. The A1 alignment records are probes *after* joint training; their mean cosine is an average of examplewise cosines, not the cosine of the aggregate parameter gradients. Their unsigned loading cannot establish the sign of $v^\top G_a$. The complete main-figure audit is generated by `paper/revision/build_figures.py`, with source hashes and reconstructed statistics. Duplicate low-label dose logs are reconciled by $(\alpha, \mathrm{mode}, \mathrm{seed})$ and checked for equality before removal.

`paper/revision/check_boundary.py` checks the exact discrepancy, the sufficient sign certificate, the realizable conflict example, and the finite-step remainder on seeded linear encoders. These are algebraic checks of the new theory, separate from the archived neural-network experiments.

A.1 Conditional Dynamical Models

Definition 2 (Degenerate minimizer)**.** *A loss has a degenerate minimizer if a constant representation attains its minimum. For a differentiable interior minimum, stationarity is necessary but insufficient. Variance shrinkage has such minima; reconstruction with varying targets generally does not, because a constant code need not be stationary for a fixed decoder. Accordingly, we evaluate trained reconstruction heads empirically.*

Theorem 2 (Conditional scalar collapse transition)**.** *Assume an independently justified scalar reduction*

$$\dot{\sigma} = -\alpha\kappa\sigma + b(\sigma), \qquad \kappa > 0, \tag{13}$$

where $b(0) = 0$. Let $f(\sigma) = b(\sigma)/\sigma$ on $(0, \infty)$ tend to zero at both ends, increase strictly to a unique nondegenerate maximum $\lambda_\star = f(\sigma_\star) > 0$, and decrease strictly afterwards. Then $\alpha_c = \lambda_\star/\kappa$. For $0 < \alpha < \alpha_c$ there are two positive equilibria, the smaller unstable and the larger stable; initial values below the smaller one still collapse. At α_c the equilibria coalesce in a saddle-node; for $\alpha > \alpha_c$, every positive trajectory tends to zero.

Proof. The vector field is $\sigma(f(\sigma) - \alpha\kappa)$. The crossings and their stability follow from monotonicity; the nonzero second derivative gives the nondegenerate fold. Above the maximum the field is strictly negative for every $\sigma > 0$, whose only possible limiting equilibrium is zero. □

Lemma 1 (Departure from a primary optimum)**.** *At $G_p = 0$, any nonzero G_a makes the joint gradient nonzero when $\alpha > 0$. This implies departure from the primary optimum, but neither collapse nor irreversibility.*

Corollary 1 (Fixed-dose attenuation)**.** *In Theorem 2, a constant auxiliary multiplier $q \in (0, 1]$ replaces the effective dose by $q\alpha$ and the threshold by α_c/q. Positive attenuation can therefore preserve the positive equilibrium at a fixed dose. Exact deletion removes the contraction term but is not necessary for fixed-dose protection.*

Theorem 3 (First modal contraction versus full linear collapse)**.** *For the linear symmetric model $\dot{z} = (R - \alpha A)z$, assume $R \succ 0$ and $A \succeq 0$ in the same coordinates. Along an eigenvector of $R - \alpha A$ with eigenvalue ν_j, the modal variance obeys*

$$\dot{\varsigma}_j^2 = 2\nu_j\varsigma_j^2. \tag{14}$$

Let

$$T = R^{-1/2}AR^{-1/2}, \qquad \rho_\star = \lambda_{\max}(T). \tag{15}$$

Some mode contracts iff $\alpha\rho_\star > 1$. Every mode contracts exponentially iff $\alpha\lambda_{\min}(T) > 1$. Equality leaves a neutral mode.

Proof. $R - \alpha A = R^{1/2}(I - \alpha T)R^{1/2}$ has the same inertia as $I - \alpha T$ by congruence. Since the generator is symmetric, its eigenvalue signs determine modal growth or decay. □

The scalar model and the symmetric linear model answer different questions. Neither follows from measuring the proxy $\hat{\rho}$ alone. For example, with $R = I$, $A = \mathrm{diag}(2, 1/2)$, and $\alpha = 1$, one mode contracts and another expands. This permits task performance on a surviving direction without preserving the full representation. The NYUv2 result is compatible with this mechanism but does not identify its dynamical eigenvectors.

B Detailed Intervention Comparisons

This section collects the per-method intervention comparisons (Table 1) and the cross-domain diagnostics (Figure 4).

Table 1: Controlled comparisons, ten seeds per method: collapse rate followed by the mean task metric. The optimizer grid uses $(m, \alpha) = (128, 0.1)$ for HAR, $(32, 3000)$ for battery, and $(64, 10000)$ for RadioML. The detached reference comes from separate R3 controls at $m = 64$ and doses 100, 100, and 3000, respectively; its auxiliary encoder update is zero.

Combiner	HAR acc. ↑	Battery log-SOH RMSE ↓	RadioML acc. ↑
JOINT (plain summation)	1.00 (0.945)	1.00 (0.611)	1.00 (0.256)
conflict resolution			
PCGrad	1.00 (0.585)	1.00 (0.601)	1.00 (0.091)
CAGrad (c=0.5)	1.00 (0.952)	1.00 (0.607)	1.00 (0.091)
Pareto			
MGDA	1.00 (0.200)	1.00 (0.385)	1.00 (0.091)
gradient geometry			
Aligned-MTL	1.00 (0.865)	1.00 (0.609)	1.00 (0.091)
IMTL-G	1.00 (0.419)	1.00 (0.601)	1.00 (0.091)
Nash-MTL	1.00 (0.280)	1.00 (0.552)	1.00 (0.091)
loss weighting			
FAMO	1.00 (0.182)	1.00 (0.613)	1.00 (0.091)
UW (uncert.)	1.00 (0.940)	1.00 (0.611)	1.00 (0.091)
GradNorm	1.00 (0.958)	1.00 (0.611)	1.00 (0.170)
IMTL-L	1.00 (0.947)	1.00 (0.611)	1.00 (0.091)
Detachment (STF-0)	0.0	0.0	0.0

Table 2: Corrected MATR test results: ten initialization seeds on a fixed cell-disjoint split, with validation-based checkpoint selection. $S = \mathrm{tr}\,\mathrm{Cov}(H)$; relative S is the mean same-seed ratio to the SOH-only reference; PR is participation rank. Censored shortfall measures lower-bound violations in cycles. Table 9 reports uncertainty.

Mode	SOH RMSE% ↓	obs. RUL MAE ↓	shortfall ↓	S	relative S	PR
SOH-only	0.61	—	—	1.01	1.00	1.35
JOINT	5.57	81.5	17.60	2.86	3.55	2.49
STF-0	**2.10**	**46.7**	**2.06**	4.77	6.24	1.86
STFCAL	2.71	63.6	5.73	1.40	1.89	2.03
STFCALRANK	4.98	107.6	17.57	617.90	651.77	1.21

A temporal encoder combines a mixer and a gated state-space block, feeding SOH, RUL, reconstruction, and distributional heads. The corrected protocol selects checkpoints on validation cells, reports observed-lifetime MAE and censored shortfall separately, and retains disjoint train/validation/test cells. Against joint training, detachment's paired mean differences across ten initializations are -3.47 SOH percentage points (95% t interval $[-4.20, -2.74]$) and -34.8 observed-RUL cycles ($[-57.8, -11.8]$). These intervals describe initialization variability on one cell split. Several protocol choices changed together; this comparison does not isolate censoring as the cause of the changed ranking. The rank arm's large covariance trace and low participation rank likewise do not establish useful diversity.

Table 3: Real auxiliary heads at width 64, ten seeds: mean primary accuracy and collapse rate in parentheses. The listed radial selective arms do not match detachment on high-dose SNR regression. The separate rank-3 arm also avoids collapse in all ten SNR seeds, but reaches only 0.211 accuracy (Figure 3c).

Auxiliary head	α	JOINT	STF-0	STF-HARD	STF-SOFT	STFCAL
HAR input-reconstruction	0.1	.942 (0.0)	.943 (0.0)	.942 (0.0)	.942 (0.0)	.942 (0.0)
HAR subject-ID probe	0.1	.941 (0.0)	.943 (0.0)	.944 (0.0)	.943 (0.0)	.942 (0.0)
RadioML I/Q-reconstruction	3000	.343 (0.0)	.307 (0.0)	.329 (0.0)	.327 (0.0)	.315 (0.0)
RadioML SNR regression	3000	.205 (1.0)	**.307 (0.0)**	.208 (1.0)	.200 (0.9)	.199 (0.1)

Table 4: Real HAR reconstruction ($m = R = 64$, ten seeds). Candidate ranks include closed-gate seeds; active ranks average only gate-open seeds. Accuracy and scatter average all ten seeds. PCA and auxiliary-gradient bases activate on identical seeds; these results establish neither superiority over joint training nor globally optimal deletion.

α	basis	gate open	candidate rank	active rank	rel. scatter	accuracy
12	PCA	4/10	57.0	62.00	0.842	0.9492
12	Aux. grad.	4/10	22.6	30.75	0.811	0.9497
30	PCA	8/10	62.8	63.62	0.754	0.9512
30	Aux. grad.	8/10	44.1	47.88	0.746	0.9531
100	PCA	10/10	64.0	64.00	0.630	0.9524
100	Aux. grad.	10/10	59.4	59.40	0.654	0.9520

The six-panel summary below retains dose sweeps, width crossings, the NYUv2 gate-boundary experiment, intervention counts, and financial outcomes. Widths change both empirical crossing locations and probe calibration (Yang & Hu, 2021); neither yields an independently identified universal capacity exponent.

The appendices cover the financial application (Appendix C), proofs and scaling analyses (Appendix D), generalizations and finite-sample theory (Appendix E), additional tables and protocols (Appendix F), NYUv2 (Appendix G), supplementary experiments (Appendix H), and matched-dose regularization baselines (Appendix I).

C Financial Ranking Application

The failure was first observed in multi-task equity selection. A production model combined a pairwise-ranking head with distributional (MDN (Bishop, 1994)), reconstruction (AE), and view-contrastive (SCARF-style (Bahri et al., 2022)) auxiliaries over a shared encoder of A-share factors. After deployment, with the training objective unchanged, predicted rankings became nearly constant. Cross-sectional prediction spreads fell by approximately 93% relative to validation, although individual training losses remained low. This observation motivates measuring representation and prediction variation alongside loss; it does not by itself identify a constant attractor or vanishing gradients.

The financial application uses a separate walk-forward model with real distributional, reconstruction, and contrastive heads. The released R5fin records set `fixed_aux=0`; Table 12 therefore does not replicate the controlled variance toxifier. Across ten folds and ten initialization seeds, mean out-of-sample (OOS) rank-IC is 0.060 for joint training, 0.036 for detachment, 0.063 for the hard gate, and 0.056 for the soft gate. After averaging seeds within each fold, paired differences from joint are -0.0237, $+0.0029$, and -0.0042, respectively. Two-sided Wilcoxon p values are 0.037, 0.432, and 0.695; none survives Holm correction over the three comparisons. Because temporal folds may remain dependent, these tests are sensitivity analyses rather than independent-population inference. The results suggest a possible cost of indiscriminate detachment without establishing retention of

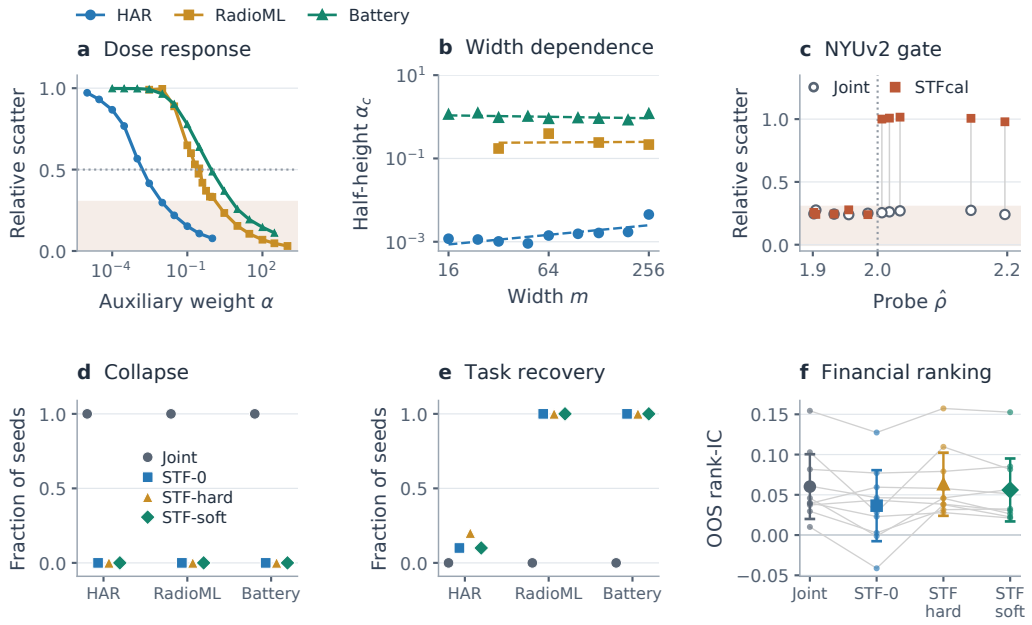

Figure 4: Additional diagnostics across domains. (a) Dose–scatter curves; the half-height crossing differs from the operational 0.3 criterion. (b) Descriptive width trends. (c) Paired NYUv2 boundary outcomes at $\alpha = 0.5$. (d,e) Collapse and recovery rates for the archived controlled interventions. (f) Financial rank-IC with temporal folds as the pairing unit. Source protocols and numerical tables are provided in this supplement.

a particular useful subspace. The script `finance_fold_audit.py` reproduces the comparisons from all 40 source files.

In the separate synthetic testbed, initialization probes give $\rho = 0.74 \pm 0.19$ over 12 seeds and predicted critical weight $\alpha_c = \langle 1/\rho \rangle = 1.45 \pm 0.37$, consistent with the observed onset. Other domains use the projected-gradient diagnostic of Definition 1, but its numerical value depends on the task, encoder, and calibration time.

D Proofs and Additional Analysis

D.1 Statements Deferred from the Main Text

We formalize several statements discussed in the main text. Theorem numbering continues from the main text.

Proposition 2 (Sufficient conditions for gradient vanishing)**.** *For a differentiable prediction map with Jacobian J and output loss gradient ℓ, the parameter gradient vanishes iff $J^\top \ell = 0$. Sufficient conditions are $\ell = 0$, or identical Jacobian rows together with $\mathbf{1}^\top \ell = 0$. Equal prediction values alone imply neither condition.*

Proof. Apply the chain rule. For a counterexample to the stronger assertion, $\hat{y}_i = \theta x_i$ is constant at $\theta = 0$ but has Jacobian rows x_i. The pairwise logistic loss $\log(1 + e^{-(\hat{y}_1 - \hat{y}_2)})$ has output gradient $(-1/2, 1/2)$ at a tie, so its parameter gradient is $(x_2 - x_1)/2$, generally nonzero. □

Proposition 3 (Scope of positive attenuation)**.** *Under Theorem 2, a fixed positive multiplier q leaves a finite threshold α_c/q but can prevent the transition at a specified dose. A dose-dependent positive multiplier can keep the effective dose subcritical for every α: for $0 < \epsilon < 1$, take $q(\alpha) = (1 - \epsilon)/(1 + \alpha/\alpha_c)$.*

Proof. Substitute $q\alpha$ for α. The proposed choice gives $q\alpha/\alpha_c = (1-\epsilon)\alpha/(\alpha+\alpha_c) < 1$. A positive stable equilibrium remains, although its lower basin boundary persists. Preserving gradient direction therefore does not imply that positive task weighting must fail. □

Proposition 4 (Rank-dilution scaling). *Assume the effective ratio scales as $\rho_\star \asymp (m_\Pi/m)^{\gamma_0}$, $\gamma_0 \geq 1$, and $m_\Pi \propto m^\beta$, $0 \leq \beta < 1$. Then $\alpha_c = 1/\rho_\star$ implies*

$$\gamma \;=\; \gamma_0\,(1-\beta), \qquad \gamma_0 \geq 1, \tag{16}$$

The separable scaling is assumed; rank alone does not imply it. More generally, restoration $\asymp m^a$ and contraction $\asymp m^b m_\Pi^c$ yield $\gamma = a - b - c\beta$, with the displayed form requiring $c = a - b = \gamma_0$. If the toxic norm scales as m_Π^p, then $c = 2p$ only when contraction depends quadratically on that norm and has no additional rank dependence. The equality $\gamma_0 = 2p$ further requires $a - b = 2p$. Neither exponent is identified by the toxic eigenvalue spectrum alone. A fixed effective rank gives $\beta = 0$; a fixed per-head output dimension need not.

Corollary 2 (Composition of rank-one PSD operators). *Let $T = \sum_k \rho_k u_k u_k^\top$, with unit u_k and $\rho_k \geq 0$, in common restoration-normalized coordinates. If the directions are mutually orthogonal, $\lambda_{\max}(T) = \max_k \rho_k$; if aligned, it is $\sum_k \rho_k$. An additional orthogonal direction leaves the maximum unchanged only if its coefficient does not exceed the previous maximum.*

Proof. For orthogonal directions, the u_k are eigenvectors. For aligned directions, the sum is a single rank-one operator. Neither case establishes the geometry of the production model's heads. □

For an explicit scalar example, take $b(\sigma) = \lambda\sigma^2/(1+\sigma^2)$. Then

$$\sigma_\pm = \frac{\lambda \pm \sqrt{\lambda^2 - 4(\alpha\kappa)^2}}{2\alpha\kappa}, \qquad \alpha_c = \lambda/(2\kappa). \tag{17}$$

This verifies the fold and its basin dependence. Radial auxiliary behavior,

$$g_a = \kappa\delta h + o(\|\delta h\|), \tag{18}$$

requires isotropic curvature; stationarity at a constant representation is insufficient. For a general encoder, the Jacobian still determines its effect on variance through (2). In discrete time, linearizing a scalar stable equilibrium (Strogatz, 1994) with derivative $-\ell$ gives

$$e_{t+1} = (1 - \eta\ell)e_t, \qquad 0 < \eta\ell < 2, \tag{19}$$

where e_t is a deviation from that equilibrium, not total representation scatter.

Power-law spectra do not determine the capacity exponent. For eigenvalues $\lambda_i = Ci^{-(1+s)}$ with C independent of rank, $\|A_r\|_F^2 = C^2 \sum_{i=1}^r i^{-2(1+s)}$. Hence the Frobenius norm is bounded for $s > -1/2$, grows as $\sqrt{\log r}$ for $s = -1/2$, and as $r^{-s-1/2}$ for $s < -1/2$. A decaying spectrum ($s > -1$) therefore has power exponent at most $1/2$; for the commonly intended $s > 0$ it has exponent zero. The trace instead involves $\sum_i i^{-(1+s)}$, and the operator norm is C for a decreasing spectrum. Rank-dependent normalization $C(r, m)$ changes these conclusions. Thus neither a universal $\gamma_0(s)$ nor $\gamma_0 = 2p$ follows without specifying the norm, normalization, restoration and coupling to dynamics. Concretely, if $\kappa \asymp m^b\|A_r\|^q$, $\|A_r\| \asymp r^p$, and $r \asymp m^\beta$, then $\gamma = a - b - qp\beta$. The restricted form $\gamma_0(1-\beta)$ requires $a - b = qp = \gamma_0$. Hence $\gamma_0 = 2p$ holds only under the additional quadratic coupling $q = 2$ and matching restoration exponent $a - b = 2p$; these are not implied by the measured width slope.

The original two-head generator has $W_k \in \mathbb{R}^{m\times r}$ with independent entries of variance $1/r$. Its combined span has rank $\min(m, 2r)$ almost surely. For the averaged quadratic loss the sample-normalized Hessian is $A = 2\sum_k W_k W_k^\top/(2r)$, with $\mathbb{E}A = 2I/r$. At $r = 32$, widths $32, 64, 128, 256$ therefore have ranks $32, 64, 64, 64$. Fixed per-head rank is not fixed combined rank over this range, and changing r also changes curvature. The supplied orthonormal control instead constructs $A = Q\,\mathrm{diag}(\lambda)Q^\top$ at exactly specified rank and normalization.

The RS1 experiment fixes per-head ranks $r \in \{4, 8, 16\}$ across widths $32, 48, 64, 96$, giving combined ranks $8, 16, 32$. With anchor 3×10^{-4}, descriptive half-height slopes are -0.31, 0.45, and 3.02. Minimum relative scatter ranges are 0.040–0.068, 0.013–0.018, and 0.0056–0.0132, respectively: some curves contract without reaching a near-zero floor. Using the unpenalized anchor changes the

slopes to 0.65, 5.55, and 3.88. This rank and anchor sensitivity does not support a single universal exponent. The released analyzer retains all twelve curves and reports crossings separately from scatter floors.

The orthonormal control constructs the sample-normalized auxiliary Hessian $A = Q\,\mathrm{diag}(\lambda)Q^\top$ at exact rank r and verifies its amplitude. Three normalizations hold different quantities fixed: $\lambda = 1$ fixes curvature per direction, so the Frobenius norm grows as $\sqrt{r}$; $\lambda = 1/r$ fixes the trace; and $\lambda = 1/\sqrt{r}$ fixes the Frobenius norm. The training grid tests the first two at $m \in \{32, 64, 128\}$, $r \in \{4, 16\}$, seven doses, and three seeds, holding the loss, optimizer, and data generator otherwise fixed. Three observations follow. First, at fixed width and per-direction curvature, the $r = 16$ crossing is lower than the $r = 4$ crossing by factors of 4.8 at $m = 64$ and 228 at $m = 32$. Second, the fitted width slope for $r = 4$ changes from -0.29 under $\lambda = 1$ to $+0.91$ under $\lambda = 1/r$. Third, minimum scatter is 3.5–11% of the healthy reference for $r = 4$, versus 0.4–1.1% for $r = 16$. Thus normalization affects the estimated exponent even when rank is controlled exactly. Each three-width fit has only one residual degree of freedom and cannot identify a precise exponent. The commands `controlled_rank_scan.py --self-test` and `run_controlled_rank_grid.py` verify the construction and reproduce the twelve configurations. The historical squared pairwise hinge has zero margin, so tied scores can attain zero primary loss. These controls concern that objective; they do not establish the same transition for informative ranking objectives. The separate `--primary-loss logistic` option removes this degeneracy, but no unfinished logistic sweep is counted as evidence.

The synthetic testbed uses random toxic projections of rank $\lfloor m/2 \rfloor$ and a low-rank primary task with 2–8 active directions across seeds (mean 5.3). It sweeps 17 logarithmically spaced doses over $[3 \times 10^{-4}, 10]$ at six widths $m \in \{8, \ldots, 256\}$ under a pairwise-ranking loss. The toxic-rank schedule is proportional to width. The relative half-height rule gives $\gamma = 0.07$ across six widths and $\gamma = 0.01$ across the four widths used below. This nearly flat trend is consistent with the restricted law at $\beta = 1$, but does not independently establish that law.

Using a fixed absolute scatter threshold instead gives $\gamma = 1.53$ ($R^2 = 0.91$) on the same curves. At $\sigma_h = 1.5 \times 10^{-3}$, the crossing depends on the width-dependent healthy amplitude: the fixed-rank family has $\sigma_h(0) \propto m^{0.95}$. The absolute rule also gives nearly constant slopes $1.94, 2.02, 2.04, 1.96, 1.89$ across rank schedules $\beta = 0, \ldots, 1$, varying by only 0.15 and largely masking schedule dependence. Using relative crossings, the fixed-per-head-rank family instead gives $\gamma_0 = 2.43$ ($R^2 = 0.90$, four widths). Its combined span can still grow as $\min(m, 2r)$, so this is not an exact fixed-effective-rank estimate.

Both synthetic families extend the dose grid to $\alpha = 0$ (21 points), using four widths $m \in \{32, 64, 128, 256\}$, five seeds, and 800 epochs. The fixed-rank family is the $\beta = 0$ case with constant per-head rank r. A corresponding ten-seed run yields crossings differing by $+3.1\%, +1.8\%, -5.8\%, +19.7\%$ across these widths. This comparison quantifies sensitivity to seed count, including variation in the healthy anchor at larger widths.

Two checks qualify the crossing estimator. First, a half-height crossing can occur without near-total collapse. For fixed per-head rank $r \leq 4$, scatter plateaus at 2–23% of its healthy value; only $r \geq 8$ reaches a floor of at most 0.9%. These floors must be distinguished from the operational 0.3 collapse threshold. Second, the synthetic fits use the first penalized dose, $\alpha = 3 \times 10^{-4}$, as the anchor. The unpenalized value is set by the ranking loss, is shared across toxic ranks, and lies 30–40% above the penalized trend. Varying the anchor over $\alpha \in [0, 3 \times 10^{-3}]$ preserves the decreasing ordering of $\gamma(\beta)$: $\gamma_0 \in [1.55, 3.23]$ at $\beta = 0$, versus $\gamma(1) \in [-0.06, +0.74]$. The ordering is more stable than the fitted coefficients.

The rank-schedule fits below test a restricted scaling hypothesis. Substituting $m_\Pi \asymp m^\beta$ into an independently established $\alpha_c \asymp (m/m_\Pi)^{\gamma_0}$ gives $\gamma = \gamma_0(1 - \beta)$; it does not establish the premise. Back-solving β from measured γ cannot serve as an independent measurement of toxic rank. The original generator audit further limits interpreting its per-head rank schedule as an effective-rank schedule.

Crossing rules affect the fitted magnitudes. For HAR, an absolute threshold $T \in [0.20, 0.50]$ gives $\gamma \in [0.66, 0.82]$, compared with 0.66 under the four-width relative rule. Battery requires a grid that reaches a healthy branch: earlier sweeps started at $\alpha = 300$, where relative scatter was already 0.07–0.11, making thresholds left-censored. Extending to $\alpha \in [10^{-4}, 300]$ with five seeds gives a

healthy branch at $\alpha \leq 10^{-3}$ (relative scatter 1.000 ± 0.005) and crossings $0.95, 0.89, 0.90, 1.18$ at $m = 32, \ldots, 256$. The fitted slope is $\gamma_{\mathrm{bat}} = 0.10$ ($R^2 = 0.43$), varying over $[-0.10, +0.12]$ across crossing rules. These four widths are consistent with weak width dependence but do not resolve a precise exponent.

Treating each width's crossing as one observation gives four-width OLS 95% intervals $[-0.07, +4.93]$ for the $\beta = 0$ schedule, $[-0.16, +1.49]$ for HAR, $[-0.24, +0.43]$ for battery, and $[-1.16, +1.20]$ for RadioML. None excludes zero. Interior schedules yield positive slopes: 2.57 (CI $[0.87, 4.27]$) at $\beta = 0.25$ and 1.86 ($[0.81, 2.91]$) at $\beta = 0.50$; $\beta = 1$ remains consistent with zero. At the reported anchor, all four remaining schedule intervals contain $\gamma_0(1 - \beta)$, whereas the $\beta = 0.75$ interval excludes it at anchor 10^{-4}. These results support schedule dependence without identifying an anchor-independent coefficient.

The denser SYN8 family supplies 26 completed (m, β) cells, ten seeds each. Since $r(64) = 32$ for every schedule, we fit $\log \alpha_c = c + \gamma_0(1 - \beta) \log(m/64)$. At anchor 3×10^{-4}, the constrained fit gives $\gamma_0 = 2.50$ (nominal OLS 95% CI $[1.94, 3.06]$). Across anchors $\{10^{-4}, 3 \times 10^{-4}\}$ and inclusion/exclusion of $m = 128$, all four conditional intervals exclude zero (lowest bound 1.04). In the unconstrained model $c + ax + b\beta x$, $x = \log(m/64)$, every interval for $a + b$ contains zero and $-b/a = 0.78$–1.03: the linear form is not rejected, not established. These are cell-level fit intervals, not independent-seed uncertainty: the schedules share seeds and the $m = 64$ curves are identical. Joint/anchor-row point estimates range from 1.35 to 2.87. At $m \geq 96$ the sampled curve never resolves the collapsed floor; half-height crossings interpolate the observed decline, not a verified collapse threshold. Thus SYN8 sharpens a conditional trend without supplying a precise physical exponent. `synth_beta_joint.py` reproduces the fits.

Nine-width HAR and battery sweeps use $m \in \{16, 24, 32, 48, 64, 96, 128, 192, 256\}$ with the same driver, seeds, epochs, and dose grids. Shared configurations agree to floating-point precision (maximum difference 2×10^{-16}). HAR gives $\gamma_{\mathrm{HAR}} = 0.38$ (95% CI $[+0.10, +0.67]$, $R^2 = 0.60$, $n = 9$); excluding the two narrowest widths gives 0.60. Battery gives $\gamma_{\mathrm{bat}} = -0.06$ (CI $[-0.17, +0.06]$), consistent with weak width dependence. The denser grid narrows conditional fit intervals but does not eliminate crossing uncertainty. For example, HAR at $m = 32$, $\alpha = 10^{-3}$ has relative scatter 0.5026 ± 0.0524, placing the half-height crossing close to a sampled grid point and making interpolation sensitive to seed variation. We report both four- and nine-width estimates rather than treating either as an identified dynamical exponent.

We also examine probe scaling across encoders. Repeating calibration with a Transformer at $m \in \{32, 64, 128\}$ and ten seeds gives probe exponents -0.93 for HAR ($R^2 = 0.99$), $+0.39$ for battery ($R^2 = 0.87$), and -0.17 for RadioML ($R^2 = 0.79$). The corresponding MLP values are -1.63 ($R^2 = 0.96$), $+0.67$ ($R^2 = 0.86$), and $+0.28$ ($R^2 = 0.40$). Signs agree for HAR and battery; RadioML is consistent with zero for both encoders, and the 95% intervals overlap in all three domains. This is evidence of broadly similar probe trends in the tested encoders, not architecture invariance. The script `probe_exponent.py` reproduces the analysis.

Probe exponents and scatter-crossing exponents estimate different quantities. Their MLP point estimates agree in sign for RadioML ($+0.28$ versus $+0.02$), but disagree for HAR (-1.63 versus $+0.38$) and battery ($+0.67$ versus -0.06). A half-height crossing locates a dose at which trained scatter reaches half its reference value; the probe estimates $1/\hat{\rho}$ at the warm-up anchor. Agreement within a factor of a few at one width (a factor of 7 for NYUv2 at $m = 64$) does not imply agreement in width scaling. Negative probe slopes do not satisfy the restricted nonnegative exponent assumed in (16), but that assumption is not a general constraint on empirical slopes. The width analysis here therefore reports crossing exponents separately, while Appendix G reports probe values as calibration statistics.

Several sweeps show bimodal per-seed scatter near an empirical crossing. This is consistent with multiple outcomes but does not uniquely identify a saddle-node bifurcation; initialization, finite training time and optimization can also generate heterogeneous outcomes.

The original GPU experiments use fixed integer seeds, bf16 autocast with TF32 matrix multiplication, and Adam (lr $= 10^{-3}$, batch 256) on an NVIDIA A100 80 GB. Exceptions are specified with their protocols. Operational collapse means scatter below 0.3 times the same-seed single-task reference; the released recovery rate compares each seed with the mean joint-training metric. Code, seeds, hyperparameters, and per-fold results are included in the repository.

E GENERALIZATIONS AND FINITE-SIZE THEORY

Proposition 5 (Fixed attenuation in a specified model)**.** *For $T \succeq 0$ in restoration-normalized coordinates, uniform attenuation $T \mapsto (1-\lambda)T$, $0 \le \lambda < 1$, multiplies the first-modal-contraction threshold by $1/(1-\lambda)$. At fixed dose the smallest λ meeting $\alpha(1-\lambda)\|T\| \le 1$ is $\max(0, 1 - 1/(\alpha\|T\|))$ when $\alpha\|T\| > 0$.*

Proof. All eigenvalues scale by $1-\lambda$; solve the displayed inequality. This statement concerns a fixed PSD model, not arbitrary signed or nonsymmetric filtered dynamics. □

Theorem 4 (Annihilation by orthogonal compression)**.** *Let $T \succeq 0$, $U = \mathrm{range}(T)$, and $Q = I - P_S$ be an orthogonal projector. Then $QTQ = 0$ iff $U \subseteq S$. Among such deletions the unique smallest deleted subspace is U, and $I - P_U$ preserves its orthogonal complement. This is an algebraic statement about two-sided compression.*

Proof. $QTQ = (T^{1/2}Q)^\top (T^{1/2}Q) = 0$ iff $\mathrm{range}(Q) \subseteq \ker T$, equivalently $U \subseteq S$. A general one-sided backward operator QT need not describe symmetric neural-network dynamics; a signed filter can even reverse contraction, so its operator norm cannot supply the missing implication. □

Corollary 3 (Subspace estimation bound)**.** *If $\|P_U - P_{\hat{U}}\| \le \epsilon$ and $Q = I - P_{\hat{U}}$, then*

$$\|QT\| \le \epsilon\|T\|, \qquad \|QTQ\| \le \epsilon^2\|T\|. \tag{20}$$

Proof. Use $T = P_U T P_U$ and $\|QP_U\| \le \epsilon$ on one or both sides. Whitening must precede this argument if T is restoration-normalized; an angle bound in other coordinates is insufficient. These are upper bounds on residual magnitude, not evidence of a nonzero residual or inevitable failure at a finite dose. □

Theorem 5 (Minimum rank for a PSD compression)**.** *Fix a PSD operator T in specified Euclidean coordinates, with eigenvalues $\kappa_1 \ge \cdots \ge \kappa_d \ge 0$, and a threshold $\tau \ge 0$. Among orthogonal projectors $Q = I - P_S$ satisfying $\|QTQ\| \le \tau$, the minimum deletion rank is*

$$r^\star = \#\{j : \kappa_j > \tau\}.$$

Deleting $S_c = \mathrm{span}\{u_j : \kappa_j > \tau\}$ attains it. Within the class of spectral deletions, S_c also minimizes any inclusion-monotone deleted energy. Uniqueness of the energy optimum requires positive energy in every additionally deleted eigendirection.

Proof. The spectral choice leaves only eigenvalues at most τ. For any $\dim S < r^\star$, dimension counting gives a unit vector $v \in S^\perp \cap S_c$. Its Rayleigh quotient satisfies $v^\top QTQv = v^\top Tv > \tau$, violating feasibility. For spectral deletions feasibility requires containing S_c, giving the energy statement by inclusion. Rotated subspaces can also attain the same minimum rank; no global minimum-energy claim over rotations follows. □

Corollary 4 (Saturation and its limits)**.** *For $\tau = 1/\alpha$, the optimal rank changes in steps as α crosses inverse positive eigenvalues and eventually equals $\dim U$. The limiting projector is $I - P_U$, which equals full detachment only on U, or globally when U is the entire gradient space. Exhausting a finite probe depth does not prove saturation of U.*

The theorem concerns a specified operator, not a mean of gradient-norm ratios. For a linear reconstruction decoder Wh with squared loss, $g_a = W^\top \delta$ (up to loss normalization) and the Hessian is $W^\top W$. The uncentered second moment relevant to normalized residual energy is

$$C_g = \mathbb{E}\left[\frac{g_a g_a^\top}{(\|g_p\| + \varepsilon)^2}\right] = W^\top \mathbb{E}\left[\frac{\delta\delta^\top}{(\|g_p\| + \varepsilon)^2}\right] W.$$

For any rank-r projector P, the squared residual is $\mathrm{tr}((I-P)C_g)$, minimized by the top r eigenvectors of C_g (the Rayleigh–Ritz principle). Its square root upper-bounds the corresponding mean norm by Jensen's inequality. It does not generally minimize that mean itself. Hessian eigenvectors coincide with this basis under isotropic weighted residual covariance; representation PCA requires its leading subspace to coincide with the leading eigenspace of C_g. Commutation gives a common

eigenbasis but does not ensure the same ordering. A nonlinear decoder additionally contributes residual-weighted second derivatives to its Hessian. Thus a nearly flat PCA residual curve is consistent with gradient energy outside leading representation components, but does not prove that the reconstruction residual is orthogonal to those components.

Table 6 reports the variance-shrinkage rank sweep. For real HAR reconstruction at $\alpha = 12$, the auxiliary-gradient and PCA bases yield mean candidate ranks 22.6 and 57.0. Both gates open in four of ten seeds, with active ranks 30.75 and 62.0 and similar aggregate accuracy (Table 4). This demonstrates basis sensitivity, not universal rank optimality. The released records also retain the three-seed RMS-basis pilot and its negative accuracy result.

In the ten-seed battery cap32 experiment, `stfcalrank` reaches its cap in every seed and gives SOH RMSE 5.26, versus 5.24 for joint training. The larger cap therefore does not establish minimum-rank protection. A single-seed full-depth audit gives representation covariance traces 3.131, 4.065, 0.906, and 0.821 for joint training, detachment, the radial calibrated arm, and the rank arm, respectively, versus 1.147 for the SOH-only reference. Corresponding normalized SOH prediction variances are 0.00826, 0.01380, 0.01253, and 0.01019, versus 0.01379. These distinct quantities establish neither representation recovery nor constant predictions. Historical checkpoints were selected on the evaluation split and remain exploratory. The corrected cell-disjoint protocol fits preprocessing on training cells and evaluates test data after validation-based checkpoint selection.

Forward and backward-only filters differ even when their Jacobians agree. For fixed U, let $\widetilde{H}_f = H - C_B HUU^\top$ and $\widetilde{H}_b = \mathrm{sg}(H) + D - C_B DUU^\top$, where $D = H - \mathrm{sg}(H)$ and sg stops gradients. Both induce the backward map $G \mapsto G - C_B GUU^\top$, but only $\widetilde{H}_b$ preserves the auxiliary input. Consequently, the loss gradients are evaluated at different forward values and need not produce the same update. Even at full rank, both retain a batch-mean gradient channel, unlike full detachment.

We isolate this distinction on HAR reconstruction at $\alpha = 100$, $m = R = 64$, with identical warm-up, seed, calibrated basis and selected rank within each pair. All ten seeds open the gate. Forward filtering gives primary accuracy 0.9507 ± 0.0033, reconstruction MSE 0.05112 ± 0.01083, and relative scatter 0.681 ± 0.039; backward-only filtering gives 0.9478 ± 0.0058, 0.02053 ± 0.00115, and 0.597 ± 0.061. Both auxiliary metrics are evaluated on unfiltered representations. The paired accuracy difference (backward minus forward) is -0.00285 (95% t interval $[-0.00815, 0.00245]$); reconstruction improves in every seed, but relative scatter decreases. Holm-adjusted Wilcoxon p values over the three outcomes are 0.242, 0.00586 and 0.0195. This exploratory comparison shows an auxiliary-fit versus scatter trade-off, not a superior primary-task algorithm. Joint and detachment accuracy are 0.9461 and 0.9431 in the same harness. A three-seed $\alpha = 12$ pilot, including its two closed-gate pairs, is also retained. The scripts `filter_semantics_audit.py` and `summarize_filter_semantics.py` supply the semantic checks, raw records and paired summary. All probe algebra runs in fp32, including the eigendecomposition outside the gradient-extraction block; a regression check verifies invariance to enclosing bf16 autocast.

A complementary experimental intervention uses $v = \nabla_\theta V$, proposed update $u_0 = G_p + \alpha G_a$, and slack $h = V - V_{\min}$. For $k > 0$ and $v \neq 0$, the closest update satisfying $v^\top u \leq kh$ is

$$u^\star = u_0 - \frac{(v^\top u_0 - kh)_+}{\|v\|^2} v.$$

The KKT conditions prove this Euclidean halfspace projection. Under a well-defined continuous flow, $\dot{h} \geq -kh$ preserves feasibility on this *fixed batch*. At $v = 0$ feasibility requires $kh \geq 0$. Discrete updates require checking the actual post-step variance: the supplied SGD implementation backtracks and restores parameters on failure. The argument covers neither unseen data nor Adam without modeling its applied update. This applies established constrained-gradient ideas (Allibhoy & Cortés, 2024), without claiming halfspace projection as novel; empirical benefits remain to be established.

We compare deletion and positive attenuation in the same UCI-HAR setup ($m = 64$, sixty epochs; Table 5). The calibrated multiplier uses the unprojected mean gradient ratio, rather than the radial diagnostic in (10), so $c = 1$ is not an identified dynamical threshold. The $c = 0.1$ setting is competitive with the fixed multipliers tested, supporting attenuation as a necessary baseline. It does not establish a universally optimal fraction of a critical dose: the constants were compared on these outcomes, seed counts are small, and some runs are non-finite. At $\alpha = 300$, seed 0 diverges for

every arm, including detachment, while auxiliary heads continue to train during detached warm-up. Such numerical failure is distinct from representation collapse.

Table 5: Matched attenuation on UCI-HAR ($m = 64$, sixty epochs). Accuracy and calibrated multiplier q are averaged over completed seeds; `div` indicates non-finite evaluation. There are three seeds per configuration, except at $\alpha = 300$, where seed 0 diverges for every arm and is excluded. Bold identifies the best fixed setting and the best calibrated setting when it exceeds the best fixed accuracy.

	accuracy			multiplier q		
arm	30	100	300	30	100	300
plain summation	.9293	`div`	`div`	1	1	1
detachment (STF-0)	.9276	.9276	`div`	0	0	0
fixed 10^{-4}	.9275	.9276	.9272	—	—	—
fixed $3{\cdot}10^{-3}$	.9278	**.9304**	.9286	—	—	—
fixed 10^{-2}	**.9301**	.9299	**.9293**	—	—	—
fixed $3{\cdot}10^{-2}$	**.9301**	.9294	.9265	—	—	—
fixed 10^{-1}	.9295	.9291	.9260	—	—	—
fixed $3{\cdot}10^{-1}$	.9242	.8842	`div`	—	—	—
calibrated $c = 0.1$	.9299	**.9306**	.9289	.076	.016	.004
calibrated $c = 0.3$	.9261	.9285	.9281	.228	.049	.012
calibrated $c = 1.0$	.9257	.9266	.9276	.759	.163	.039

Table 6: Selected rank versus auxiliary dose on UCI-HAR ($m = 64$, five seeds, probe depth $R = 64$). Ranks increase with dose; full detachment blocks all 64 feature directions. A closed gate removes no directions. Relative scatter is $\mathcal{S}_h/\mathcal{S}_{single}$ for the representation-PCA rank arm (`stfcalrank`); candidate ranks are reported for both bases. These empirical ranks do not establish the PSD optimum in Theorem 5.

α	gate	$r^\star$ (principal)	$r^\star$ (aux. basis)	rel. scatter (rank arm)	rel. scatter (JOINT)
0.003	0	–	–	0.404	0.402
0.01	0	–	–	0.287	0.280
0.03	1	1.6	1.6	1.007	0.200
0.1	1	3.8	3.8	1.026	0.138
0.3	1	4.2	4.2	1.018	0.101
1.0	1	11.6	11.6	0.978	0.070
3.0	1	25.8	25.8	0.987	0.051
10	1	39.8	40.0	0.986	0.035
30	1	47.4	47.8	0.968	0.025
100	1	61.6	61.0	0.950	0.016

Proposition 6 (Sampling fluctuation of a frozen statistic)**.** *For independent identically distributed scalar diagnostics X_i with finite variance σ_X^2, $\hat{\rho} = B^{-1}\sum_i X_i$ has variance σ_X^2/B and obeys a central limit theorem. Under a normal approximation the probability that the* gate $\alpha\hat{\rho} > 1$ *opens is* $\Phi(\sqrt{B}(\rho - 1/\alpha)/\sigma_X)$*. This is not the probability of training collapse. Batch-dependent centering or fitted bases require additional analysis or an independent frozen calibration sample.*

Proof. Use variance additivity and the central limit theorem. For a ratio A/P, the delta-method variance includes $-2\mu_A \operatorname{Cov}(A, P)/\mu_P^3$ as well as the two marginal variance terms. Mean per-example ratios and ratios of means are distinct estimands. □

Proposition 7 (Conditional calibration bound)**.** *Assume independent $X_i \in [0, M]$ with common mean ρ and population variance at most v_0. Bernstein's inequality gives*

$$\mathbb{P}(|\hat{\rho} - \rho| \geq \varepsilon) \leq 2\exp\left[-\frac{B\varepsilon^2}{2v_0 + 2M\varepsilon/3}\right]. \tag{21}$$

Thus an error probability at most δ follows if

$$B \geq (2v_0 + 2M\varepsilon/3)\varepsilon^{-2}\log(2/\delta). \tag{22}$$

Table 7: Historical battery results with evaluation-based checkpoint selection. SOH suggests a benefit from detachment, but these are not independent test estimates. The legacy RUL score averages error against observed lifetimes and censored lower bounds; it is not observed-only lifetime MAE. Single-task rows are references, not performance ceilings. Reference rows use four seeds; other rows use ten.

	MATR		NASA	
Mode	SOH RMSE (%)	legacy RUL score	SOH RMSE (%)	legacy RUL score
single-task reference	0.51	694.6	9.17	18.5
JOINT (aux heads)	5.24 ± 0.54	209.0 ± 59.8	11.42 ± 0.85	17.5 ± 1.1
JOINT (toxifier)	1.23 ± 0.44	306.6 ± 48.2	9.66 ± 0.44	17.6 ± 2.4
STF-0	1.68 ± 0.41	574.7 ± 80.7	9.35 ± 0.55	18.0 ± 2.0
STF-0W (tri-weight)	1.54 ± 0.44	541.2 ± 91.3	9.40 ± 0.62	18.6 ± 1.4
STFCAL ($\hat{\rho}$=17.4/0.40)	1.77 ± 0.47	419.8 ± 33.6	11.40 ± 0.78	18.3 ± 1.7

All MATR rows: censoring-aware RUL supervision, all 185 cells retained; bf16 autocast training, fp32 evaluation.

A nonzero primary-gradient denominator alone does not provide a uniform bound M. Population bounds cannot be replaced by observed sample moments without a separate valid empirical-concentration result.

Proof. Apply the bounded-variable Bernstein inequality to the centered independent sum and solve for B. □

Corollary 5 (Certifying a frozen gate or nested rank rule)**.** *Under Proposition 7, a gate with population margin $d = |\rho - 1/\alpha| > 0$ has error probability at most $2\exp[-Bd^2/(2v_0 + 2Md/3)]$. For $R+1$ fixed residual statistics with uniformly positive margins, a union bound controls all their threshold decisions and hence the selected nested rank. This does not identify the statistics with Hessian eigenvalues, certify an adaptively fitted basis without conditioning, or certify a subsequent training trajectory.*

Proof. A wrong threshold decision requires an estimation error at least its margin; apply the bound and then the union bound. Strict margins exclude equality cases. □

Proposition 8 (Local linear noise floor)**.** *For deviations $e_{t+1} = ae_t + \eta\xi_t$ from a fixed equilibrium, assume $|a| < 1$ and independent mean-zero innovations of variance D. The stationary variance is $\eta^2 D/(1-a^2)$.*

Proof. The second-moment recursion is $\mathbb{E}e_{t+1}^2 = a^2\mathbb{E}e_t^2 + \eta^2 D$; sum the convergent geometric series. This is variance of deviations, not total representation scatter. Divergence of a local linear approximation as $|a| \to 1$ does not prove a universal ordering of half-height and binary-collapse thresholds in a nonlinear experiment. □

Repeated probes at a frozen HAR anchor assess sampling variation. Across 2000 resampled batches per size and five seeds, fitted slopes are -0.500 for the coefficient of variation and -0.494 for standard deviation; the mean is approximately stable (slope $+0.006$). Table 8 reports plug-in envelopes using estimated variance and range. Because these moments are estimated from the same data and examples share a batch center, the envelopes are diagnostic comparisons rather than population certificates. The probe also distinguishes mean per-example ratios from ratios based on batch-mean gradients; the latter can vanish for centered variance losses. The two estimands are not interchangeable.

E.1 Prediction Variance Normalization and Target z-Scoring

A linear prediction head $\hat{y} = w^\top h + b$ introduces a scale ambiguity: shrinking centered representations by $1/c$ and expanding w by c preserves predictions. Shrinking both generally does not preserve the regression loss, nor can constant predictions fit varying targets exactly. Consequently prediction

Table 8: Frozen-anchor sampling experiment on HAR ($m = 64$, four detached warm-up epochs, five seeds). Each seed uses 2000 batches of size B, sampled with replacement. Standard deviation and CV describe variation across batches at a fixed anchor. "Meas." is the observed frequency of $|\hat{\rho} - \rho| > \eta\rho$; "bound" is the Bernstein plug-in envelope from (21), using estimated probe moments ($\kappa^2 = 2\rho^2 c_0^2 + \frac{2}{3} R_{\max}\varepsilon$). Values above one are uninformative probability bounds. The envelopes exceed observed frequencies but are not population confidence guarantees. The fitted CV slope is -0.500 ($R^2 = 0.96$), and the mean slope is $+0.006$. The norm of $\sum_i g_a^{(i)}$ is at most 4×10^{-8} of $\sum_i \|g_a^{(i)}\|$. Raising nominal confidence from 90% to 99.9% multiplies the sample-size expression by $\ln(2/10^{-3})/\ln(2/0.1) = 2.54$.

				$\eta = 0.2$		$\eta = 0.1$	
B	$\hat{\rho}$	sd($\hat{\rho}$)	CV($\hat{\rho}$)	meas.	bound	meas.	bound
16	74.30	69.52	0.947	0.820	1.94	0.911	1.98
32	76.69	50.40	0.664	0.740	1.90	0.866	1.97
64	77.03	36.00	0.471	0.657	1.83	0.836	1.94
128	77.00	25.34	0.332	0.526	1.71	0.756	1.90
256	77.74	18.24	0.237	0.378	1.49	0.657	1.83
512	77.71	12.92	0.168	0.221	1.15	0.538	1.69
1024	77.97	9.20	0.119	0.094	0.70	0.384	1.45
2048	78.00	6.34	0.082	0.022	0.29	0.218	1.07
4096	77.90	4.54	0.059	0.003	0.06	0.095	0.61

variance and across-example representation covariance should both accompany task error. Prediction normalization and target scaling are practical controls, not proofs of a universal prediction-level trap.

The historical regression recipe standardizes targets and divides predictions by a detached batch standard deviation. This changes optimization and can reduce sensitivity to scale, but detached normalization does not implement the derivative of a truly scale-invariant loss. Softmax also does not normalize logit scale: changing that scale changes probabilities. These operations therefore require empirical ablations, with scaling statistics fit only on training data.

F Additional Tables and Protocol

The corrected battery protocol uses cell-disjoint training/validation/test partitions: 130/27/28 cells for MATR and 24/4/5 for NASA. MATR test data contain 1069 observed and 1797 censored cycles; all 344 NASA test cycles are observed. Reported variation is over initializations on these fixed splits. Checkpoints minimize validation SOH, prioritizing that task rather than exploring a two-task Pareto frontier. Participation rank is $(\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$; prediction variance is measured in squared SOH percentage points. A large covariance trace with low participation rank does not establish a rich representation, and neither statistic is invariant to arbitrary encoder–head reparameterization. The script `current_evidence_audit.py` checks split disjointness and seed pairing, records source hashes, and regenerates the summary and table. This section also collects the dataset summary (Table 11), the historical censoring ablation (Table 13), the battery-head comparison (Table 14), the auxiliary-weight sweep (Table 15), and the training protocol (Algorithm 2).

On NASA, the corrected protocol gives SOH RMSE 14.43 for joint and 12.74 for detachment, against the SOH-only reference 12.66. The paired difference is -1.69 percentage points (95% t interval $[-2.54, -0.83]$, nine of ten seeds improve). Observed-RUL MAE changes from 20.05 to 19.43 cycles, but the paired interval $[-3.37, 2.13]$ includes zero (exact two-sided Wilcoxon $p = 0.770$). The calibrated gate stays closed in all ten seeds: its two arms coincide with each other (14.47 SOH, 19.20 RUL), but not with joint because their encoder warm-up is detached. NASA has only five test cells; initialization intervals cannot establish population generalization.

In the historical ten-seed MATR experiment, the radial calibrated arm improves SOH over joint training (1.77 ± 0.47 versus 5.24 ± 0.54), while detachment reaches 1.68 ± 0.41. The cap32 rank arm remains near joint (5.26), so the radial result does not validate minimum-rank protection; Corollary 4 locates the saturation limit at full-width deletion. Censored RUL requires separate observed-error

Table 9: Corrected battery test results. Means ± population standard deviations over ten initialization seeds on one fixed cell split. RUL is observed-only MAE; S is across-example covariance trace, PR is participation rank, and V_{pred} is SOH prediction variance in squared percentage points. The SOH-only head does not train RUL.

Data	Mode	SOH RMSE (%)	RUL MAE	S	PR	V_{pred}
MATR	SOH-only	0.61 ± 0.19	—	1.01 ± 0.50	1.35 ± 0.14	141.67 ± 5.92
MATR	JOINT	5.57 ± 0.48	81.54 ± 26.84	2.86 ± 0.57	2.49 ± 0.65	66.10 ± 8.59
MATR	STF-0	2.10 ± 0.62	46.73 ± 11.32	4.77 ± 1.33	1.86 ± 0.22	109.50 ± 13.01
MATR	STFCAL	2.71 ± 0.54	63.59 ± 19.43	1.40 ± 0.44	2.03 ± 0.20	99.26 ± 9.69
MATR	STFCALRANK	4.98 ± 1.09	107.58 ± 31.10	617.90 ± 489.76	1.21 ± 0.25	70.11 ± 20.06
NASA	SOH-only	12.66 ± 0.99	—	3.39 ± 1.57	2.59 ± 0.74	183.51 ± 44.46
NASA	JOINT	14.43 ± 0.64	20.05 ± 1.59	5.41 ± 0.85	3.25 ± 0.49	127.49 ± 24.32
NASA	STF-0	12.74 ± 1.20	19.43 ± 2.66	4.38 ± 1.28	3.16 ± 0.58	135.70 ± 37.07
NASA	STFCAL	14.47 ± 1.22	19.20 ± 2.40	4.55 ± 1.42	2.75 ± 0.62	109.72 ± 24.59
NASA	STFCALRANK	14.47 ± 1.22	19.20 ± 2.40	4.55 ± 1.42	2.75 ± 0.62	109.72 ± 24.59

Algorithm 2 STF training protocol (per domain, per configuration)

1: **Input:** domain data $\mathcal{D}$, encoder width m, toxicity schedule $\{\alpha\}$, seeds
2: **Phase 1 – health anchor:** train the single-task encoder ($\alpha = 0$), record healthy scatter $\mathcal{S}_{single} = \mathcal{S}_h(\theta^*)$ and healthy spectral ratio ρ_0.
3: **Phase 2 – calibration:** sweep α with mode JOINT; record ρ, relative scatter $\mathcal{S}_h/\mathcal{S}_{single}$, collapse flag; locate α_c by the threshold crossing.
4: **Phase 3 – repair:** at a selected dose, train each arm on $L_p(H) + \alpha L_a(\widetilde{H})$, using its specified forward map $\widetilde{H}$ (or detached H for STF-0); backpropagate this loss and monitor the gradient diagnostic and relative scatter.
5: **Report:** collapse rate, recovery rate, ρ, $\mathcal{S}_h/\mathcal{S}_{single}$.

and lower-bound-shortfall metrics; a lower legacy MAE need not indicate better lifetime prediction. NASA's closed gate and its discrepancy from detachment further limit the diagnostic. Because these historical checkpoints were evaluation-selected, the corrected validation protocol provides the primary battery evidence.

Algorithm 3 Released calibrated forward filters: radial STFCAL and capped STFCALRANK

1: **Input:** data $\mathcal{D}$, encoder width m, operating weight α, probe depth R, warm-up length W
2: **Warm-up:** train θ with the auxiliary branch detached ($\nabla_\theta L_a \equiv 0$) for W epochs; record the healthy anchor $\mathcal{S}^\star$ of (8)
3: **Probe:** obtain per-example gradients on an unfiltered frozen batch, radial projectors $\Pi_{c,i}$, and principal basis U. With $\varepsilon = 10^{-8}$, compute $\hat{\rho} = B^{-1}\sum_i \|\Pi_{c,i} g_a^{(i)}\|/(\|g_p^{(i)}\| + \varepsilon)$ and $\hat{\rho}_{\text{res}}[r] = B^{-1}\sum_i \|(I - U_r U_r^\top) g_a^{(i)}\|/(\|g_p^{(i)}\| + \varepsilon)$ for $r = 0, \ldots, R$.
4: **Calibrate:** $g \leftarrow \mathbf{1}[\alpha\hat{\rho} \geq 1]$; $\mathcal{R} \leftarrow \{r \leq R : \alpha\hat{\rho}_{\text{res}}[r] < 1\}$; choose $r \leftarrow \max(1, \min \mathcal{R})$ if nonempty, else $r \leftarrow R$. Record whether a feasible rank exists.
5: **Latch:** freeze g, r, U. If $g = 0$, pass H to the auxiliary head. Otherwise, STFCALRANK passes $H - C_B H U_r U_r^\top$; STFCAL passes $h_i - \langle h_i - \bar{h}, u_i\rangle u_i$, $u_i = (h_i - \bar{h})/\max(\|h_i - \bar{h}\|, \varepsilon)$. Differentiate the resulting auxiliary loss through this map.
6: **Report:** gate, r, $\hat{\rho}$, $\mathcal{S}_h/\mathcal{S}^\star$

A four-seed MATR ablation removes each weighting factor in turn. Omitting the rate factor gives fast-charge SOH RMSE 0.95; omitting the temperature factor gives extreme-temperature SOH RMSE 2.35. These results support the intended association between factors and operating regimes. A sweep over $\omega \in \{0.5, 1, 2\}$ evaluates sensitivity around the chosen value $\omega = 1$.

Table 21 reports the composition experiment. On HAR, accuracy changes from 0.943 with filtering alone to 0.933 with PCGrad and 0.940 with CAGrad. Relative scatter is at most 0.15 without filtering, compared with 0.97–1.14 after filtering. These measurements support compatibility for the tested filter-then-combiner order.

Table 10: The sixteen controlled toxifier cells behind Figure 2a. Every training auxiliary is a fixed projected-variance toxifier; the probe uses an isotropic variance surrogate. $\alpha_{\mathrm{probe}} = 1/\bar{\rho}$ is the inverse mean warm-up probe, while $\alpha_{1/2}$ interpolates the mean dose–scatter curve at relative scatter 0.5. Their ratio is a calibration-transfer diagnostic. Each probe value is a ten-seed mean of the archived warm-up records.

Domain	Head	Width m	Probe $\bar{\rho}$	α_{probe}	$\alpha_{1/2}$	Ratio
HAR	toxifier	32	30.2	3.32×10^{-2}	1.02×10^{-3}	32.5
HAR	toxifier	64	110.2	9.07×10^{-3}	1.41×10^{-3}	6.4
HAR	toxifier	128	552.0	1.81×10^{-3}	1.63×10^{-3}	1.1
HAR	toxifier	256	1504.9	6.65×10^{-4}	4.52×10^{-3}	0.15
Battery	toxifier	32	1.10	9.07×10^{-1}	9.51×10^{-1}	0.95
Battery	toxifier	64	0.71	1.42	8.95×10^{-1}	1.58
Battery	toxifier	128	0.85	1.18	9.03×10^{-1}	1.31
Battery	toxifier	256	0.28	3.56	1.18	3.01
RadioML	toxifier	32	0.14	7.13	1.73×10^{-1}	41.1
RadioML	toxifier	64	0.08	11.9	3.95×10^{-1}	30.3
RadioML	toxifier	128	0.05	19.8	2.41×10^{-1}	82.2
RadioML	toxifier	256	0.04	24.9	2.16×10^{-1}	115.6
NYUv2	toxifier	32	2.26	4.43×10^{-1}	1.14×10^{-1}	3.9
NYUv2	toxifier	64	2.01	4.98×10^{-1}	6.89×10^{-2}	7.2
NYUv2	toxifier	128	1.64	6.11×10^{-1}	3.56×10^{-2}	17.1
NYUv2	toxifier	256	1.29	7.78×10^{-1}	2.39×10^{-2}	32.6

Table 11: Four scientific domains and the financial discovery application. Sample counts describe the listed experimental datasets; individual protocols specify their partitions. Controlled scientific experiments use variance-shrinkage auxiliaries. The financial application uses real heads, and the battery and NYUv2 application studies also evaluate practical reconstruction heads.

Domain	Signal	Primary task	Samples	Metric
HAR (UCI-HAR)	motion (561-d)	6-class activity	7,352 / 2,947	accuracy
RadioML (2016.10A)	I/Q (256-d)	11-class modulation	110,000 / 110,000	accuracy
Battery (NASA+MATR)	charge/discharge	SOH + RUL	18,696 / 3,210	RMSE / MAE
Finance (A-share)	1,680-d factors	5-d return rank	800 stocks × 2,059 d	rank-IC
NYUv2	RGB-D (144×192)	13-class seg + depth + normals	795 / 654 scenes	pxacc / mIoU / RMSE / angle

The synthetic testbed uses 17 log-spaced doses, six widths, and random projections of rank $\lfloor m/2 \rfloor$. Its near-flat relative-crossing slope ($\gamma = 0.07$) contrasts with the absolute-threshold estimate (1.53), as analyzed above. In RadioML, a five-seed sweep over $\alpha \in [3\times10^{-3}, 1]$, refined near the transition, gives crossings $\{0.17, 0.39, 0.24, 0.22\}$ at widths $\{32, 64, 128, 256\}$. The eightfold width range changes the crossing by at most a factor of 2.3, much less than the factor of eight implied by $\gamma = 1$. At $m = 64$, per-seed crossings span 0.29–0.54. Together with the near-zero fit R^2, this variation limits inference about the width slope.

The repository supplies the scripts and recorded results listed in its reproduction manifest; data-dependent reruns additionally require the stated datasets and preprocessing. New audits are identified separately from historical experiments. The drivers `cross_domain_mtl.py`, `battery_bmtl_v3.py`, and `nyu_bmtl.py` implement the STF family for the unified, battery, and dense-perception experiments. Algorithm 3 specifies the calibrated forward maps; each driver records application-specific details. PCGrad, CAGrad, and FAMO share these experiment pipelines, including the composition study. The remaining methods (MGDA, Aligned-MTL, Nash-MTL, IMTL-G, IMTL-L, GradNorm, and uncertainty weighting) use `moo_combiners.py`, which ports

Table 12: Financial application: mean OOS rank-IC ± population s.d. over 100 fold-seed cells (ten temporal folds, ten initialization seeds). Cells sharing a fold are dependent. After averaging seeds within folds and correcting three paired comparisons, no method differs significantly from joint training. These records use real auxiliary heads without the fixed variance toxifier.

Mode	OOS rank-IC	vs. joint
JOINT	$+0.060 \pm 0.055$	—
STF-0 (`.detach()`)	$+0.036 \pm 0.056$	-39%
STF-HARD	$+0.063 \pm 0.052$	$+5\%$
STF-SOFT	$+0.056 \pm 0.053$	-7%

Table 13: Historical MATR censoring ablation: SOH RMSE (%) and legacy RUL MAE (cycles), with leave-cell-out evaluation, $d = 128$, bf16 training, and fp32 evaluation. Legacy rows use four seeds; the censoring-aware row uses ten. Legacy supervision treats censored bounds as exact targets; the EOL-only variant additionally excludes $96/185$ cells, leaving 14 test cells. Censoring-aware supervision applies a one-sided hinge to the 59.2% censored samples and retains all cells. Evaluation-selected checkpoints and legacy RUL scoring limit interpretation.

	SOH RMSE (%)		RUL MAE (cyc)		
Mode	JOINT	STF-0	JOINT	STF-0	cells used
legacy (RUL=0 labels)	5.56 ± 0.17	2.05 ± 0.38	94.1 ± 5.7	97.0 ± 15.3	185
legacy + `--eol-only`	1.53 ± 0.13	0.53 ± 0.03	77.5 ± 2.3	76.6 ± 4.1	89
censoring-aware (ours)	5.24 ± 0.54	1.68 ± 0.41	209.0 ± 59.8	574.7 ± 80.7	185

The legacy rows reproduce the `nocens` supervision, every other setting held fixed.

kernels from the authors' reference implementations. Comparisons use the recorded optimizer and schedule with published method defaults; batch-size exceptions are stated below.

JSON job manifests specify each experimental configuration by domain, width, mode, and seed, together with its command, output path, and log path. A deterministic runner skips existing outputs. The manifests cover collapse scans, width sweeps, filter composition, real auxiliary heads, battery weight sweeps, calibrated validation, NYUv2 experiments, and the three-domain multi-task baseline comparison. Experiments use fixed integer seeds, with counts reported for each study. Where available, JSON records retain individual seed outcomes; aggregate-only historical records are identified in the artifact audit.

The original unified runs use an NVIDIA A100 80 GB, Adam ($\mathrm{lr} = 10^{-3}$), bf16 autocast, and TF32 matrix multiplication. Calibration runs in fp32 outside autocast on detached features. Batch size is 256 for the unified protocol, real-head comparison, regularizer sweep, and MATR application; 128 for NASA conflict baselines versus 256 for its plain-summation runs; 32 for NYUv2; and 512 for financial walk-forward training. The NASA batch-size difference qualifies those historical comparisons. A unified-protocol configuration takes a few GPU-minutes; the battery application uses 150 epochs per configuration. Corrected battery runs use Adam at 2×10^{-3}, the parameter-constraint pilot uses SGD, and exact-rank controls use AdamW on CPU. Individual scripts and JSON configurations record these optimizer differences. Calibration records contain each seed's $\hat{\rho}$, residual statistics $\hat{\rho}_{\mathrm{res}}[0..R]$, gate, and selected rank. The script `calib_cert_b.py` repeats probes on a frozen healthy anchor at different batch sizes to produce Table 8.

Table 16 measures per-epoch training cost on UCI-HAR at four widths. We report the median of per-repetition median epoch times over ten repetitions, exclude epoch 0, and synchronize the device within each timed epoch. Runs begin after three consecutive idle-device readings, and method order rotates between repetitions to reduce timing drift. Contention materially affects the result: a non-idle run measured the reference at $0.180\,\mathrm{s}$ rather than $0.099\,\mathrm{s}$ and FAMO overhead at 0.0% rather than approximately 31%.

On the idle device, hard and soft filtering cost at most 21.8% on HAR. At $m = 256$, their overheads are 9.5% and 12.2%, respectively. For the tested radial implementation, projection costs $O(Bd)$ compared with the encoder's $O(Bd^2)$; this accounting does not include arbitrary full-rank eigendecompositions. Detachment and calibrated radial filtering reduce measured time by up to 65.0% and

Table 14: Historical battery-head comparison at $\alpha = 10$ (five seeds): SOH RMSE (%), legacy all-sample RUL MAE (cycles), and "scatter" $\mathbb{E}_b[\mathrm{Var}_j(h_j^{(b)})]$ (sample correction $d - 1$, no centering across examples), not (8). PCGrad, FAMO, and CAGrad yield higher SOH error than detachment in both datasets. Scatter alone does not establish operational collapse; Table 1 reports separate controlled-toxifier experiments. These evaluation-selected results are exploratory.

	MATR			NASA		
Combiner	SOH RMSE (%)	RUL MAE	scatter	SOH RMSE (%)	RUL MAE	scatter
single-task reference	0.51	694.6	–	9.17	18.5	–
JOINT (sum, plain)	6.43 ± 0.19	208.8	0.126	13.33 ± 1.07	18.7	0.214
PCGrad	4.02 ± 0.98	172.9	0.094	11.89 ± 0.73	18.6	0.164
FAMO	10.10 ± 2.17	295.3	0.435	16.02 ± 0.50	32.4	0.474
CAGrad	6.33 ± 0.45	198.0	0.166	13.04 ± 0.75	17.7	0.198
STF-0 (ours)	1.83 ± 0.51	623.4	0.198	8.81 ± 0.72	17.1	0.225

Table 15: Historical MATR auxiliary-weight sweep (seed 0): SOH RMSE (%) and legacy RUL MAE (cycles), with censoring-aware supervision and bf16 training/fp32 evaluation. Joint training uses reconstruction/distributional weight λ_{aux}; detachment makes encoder learning independent of this weight. SOH error generally increases with dose, although the $d = 32$ curve is not strictly monotone. Single-task rows are width-matched references; legacy RUL scores and evaluation-selected checkpoints are not independent lifetime-error estimates.

		SOH RMSE (%)↓			RUL MAE (cyc)↓		
Mode	λ_{aux}	d=32	d=128	d=256	d=32	d=128	d=256
JOINT	10^{-2}	2.72	1.36	1.52	478.0	386.3	377.5
	10^{-1}	2.69	2.23	2.87	230.4	399.7	411.0
	1	5.96	4.69	5.53	285.6	198.0	176.4
	10	6.52	6.58	6.01	178.5	197.0	182.0
STF-0 (dose-free)	—	2.36	1.18	1.51	607.2	627.4	540.2
single-task reference	—	0.60	0.33	0.57	951.8	761.7	325.8

47.8%. PCGrad, CAGrad, and FAMO cost 21.1–41.6%, 38.5–75.2%, and 19.6–31.1%, respectively. Across five domains and four widths, spectral-variant overhead is 6.3–26.0%, versus 5.8–98.2% for the comparison methods. Hard and soft filtering are cheaper than PCGrad and CAGrad in every measured configuration; FAMO is an exception to the broader ordering on CelebA and NYUv2 at $m = 256$. These are empirical timings for the specified implementations.

G NYUv2: Dense Perception Beyond Time Series

NYUv2 (Silberman et al., 2012; Ladický et al., 2014) contains 795 training and 654 test scenes with semantic segmentation (13 classes), depth, and surface normals. Images are resized to 144×192; segmentation pixels labeled -1 and zero-norm normal targets are masked. The convolutional encoder uses four stride-2 blocks with batch normalization and ReLU ($3 \to 32 \to 64 \to 128 \to 256$ channels), followed by a $d = 128$ representation and linear task heads. Losses are cross-entropy, L_1 depth error, and cosine normal error. We report final-epoch pixel accuracy, mIoU, depth RMSE/MAE, and mean angular error over ten seeds, without test-based checkpoint selection. The *proj* auxiliary is a fixed random variance-shrinkage projection to $d/4$, evaluated at $\alpha = 1$. The *ae* auxiliary is a trainable three-layer input-reconstruction head, evaluated at $\alpha = 250$, or 3.4 times its probe-based threshold. A separate unified-protocol experiment uses an MLP and a 9×12 coarse depth grid (108 outputs), with the optimizer, probe, and collapse criterion of Section 4.1.

After four detached warm-up epochs, the unified MLP probe gives $\hat{\rho} = 2.26 \pm 0.16$, 2.01 ± 0.10, 1.64 ± 0.09, and 1.29 ± 0.05 at widths $32, 64, 128, 256$. Per-seed ranges are 2.06–2.65, 1.90–2.20, 1.51–1.77, and 1.22–1.38. The inverse-mean calibration values are 0.44, 0.50, 0.61, and 0.78, yielding a probe-based width slope of 0.27 ($R^2 = 0.98$, 95% CI $[+0.15, +0.40]$).

Table 16: UCI-HAR per-epoch overhead relative to plain summation (batch 256). Entries are medians over ten repetitions of median epoch time; epoch 0 is excluded and each epoch includes device synchronization. Method order rotates, and timing begins after three idle-device readings. A within-rotation paired-ratio estimator differs by at most five percentage points. Absolute reference times are shown in the first row. Negative overhead denotes a measured speedup; the table does not measure the cost of every possible probe or rank configuration.

Method	m=32	m=64	m=128	m=256
plain summation (s/epoch)	0.0813	0.0986	0.1283	0.1454
overhead relative to plain summation				
STF-0 (detach)	−35.8%	−48.1%	−59.5%	−65.0%
STF-HARD	+21.8%	+18.1%	+14.3%	+9.5%
STF-SOFT	+21.8%	+19.8%	+8.9%	+12.2%
STFCAL	−0.7%	−20.1%	−37.7%	−47.8%
PCGrad	+41.6%	+31.4%	+26.6%	+21.1%
CAGrad	+75.2%	+60.2%	+46.0%	+38.5%
FAMO	+31.0%	+31.1%	+23.9%	+19.6%

This is not the scatter-crossing exponent. Half-height crossings on the same widths yield $\gamma = -0.77$; varying the crossing level from 0.3 to 0.7 gives $[-0.94, -0.73]$ with $R^2 \geq 0.99$, whereas HAR remains positive ($+0.34$ to $+0.46$). At $m = 64$, the crossing and calibration values are 0.069 and 0.50, a factor of 7 apart. Extending the grid four decades lower restores relative scatter to 0.99–1.01 at $\alpha = 10^{-6}$, compared with 0.64–0.83 at the previous lowest dose 0.01, without materially changing the crossing. The opposite width trends therefore cannot be explained solely by the original grid's lower endpoint.

We report the probe because it determines the implemented gate, while retaining the crossing estimate as a separate diagnostic. The probe slope varies over a range of 0.17 across calibration times on NYUv2, versus 2.55 on HAR (Appendix H.3); this relative stability does not make it a capacity exponent. For the convolutional encoder, the *proj* probe gives $\hat{\rho} = 3.27 \pm 0.54$ (range 2.45–3.92), predicting approximately 0.31, while the *ae* probe gives 0.0136 ± 0.0015, predicting approximately 73.

Table 17 shows joint collapse in $10/10$ seeds for every tested $\alpha \geq 0.5$. Calibrated filtering achieves $0/10$ only for $\alpha \geq 1$, when all gates open and relative scatter is approximately 1.00; the recorded candidate rank is 3. At $\alpha = 0.5$, the gate opens in the five seeds with $\hat{\rho} > 2$, and all five avoid collapse (candidate rank 1). This is an empirical threshold comparison, subject to Corollary 5's scope restrictions. At lower doses, closed-gate outcomes are close to joint training.

The task metric can conceal shrinkage: joint depth RMSE falls overall from 1.335 to 1.092 (approximately 18%) while relative scatter falls from 0.62 to 0.04. Thus the high-dose joint model has lower depth error than the filtered models despite meeting the representation-collapse criterion. At every tested $\alpha \geq 1$, paired Wilcoxon tests on relative scatter give $p \approx 2 \times 10^{-3}$, the smallest two-sided value attainable with ten nonzero, same-sign paired differences.

Table 18 compares dense tasks within a shared small-backbone protocol, without claiming competitive absolute NYUv2 performance. Its driver records mean within-example feature variance, not the unified statistic in (8). We therefore interpret this table as a task and feature-amplitude comparison, not evidence of across-example representation collapse. With the projection auxiliary, the reported amplitude ratio is 0.22, segmentation mIoU decreases, and depth RMSE improves; radial filtering returns the amplitude near the auxiliary-free reference. With reconstruction at $\alpha = 250$, filtered amplitude is 3.84, versus 3.99 for joint training and 7.19 for detachment. These values do not establish representation recovery. An across-example covariance audit is needed before attributing the dense-model behavior to collapse or basis drift.

Table 17: NYUv2 under the unified protocol (MLP, $m = 64$, ten seeds): depth RMSE in meters on a 9×12 grid and scatter relative to the same-seed single-task anchor. Calibration gives mean $\hat{\rho} = 2.01$ and inverse-mean threshold 0.50. Recorded candidate ranks are 1 at $\alpha \leq 0.5$ and 3 at $\alpha \geq 1$; the radial STFCAL arm does not use this rank. Full detachment yields RMSE 1.353 and relative scatter 1.000 at every dose.

α	0.03	0.1	0.3	0.5	1	3	10	30	100
JOINT RMSE	1.335	1.336	1.316	1.307	1.305	1.280	1.221	1.100	1.092
JOINT rel. scatter	0.62	0.45	0.30	0.26	0.22	0.16	0.10	0.04	0.04
JOINT collapsed	0/10	0/10	5/10	10/10	10/10	10/10	10/10	10/10	10/10
STFCAL RMSE	1.339	1.327	1.314	1.328	1.354	1.353	1.353	1.354	1.354
STFCAL rel. scatter	0.63	0.45	0.31	0.63	1.01	1.01	1.01	1.00	1.01
STFCAL collapsed	0/10	0/10	4/10	5/10	0/10	0/10	0/10	0/10	0/10
calibrated gate open	0/10	0/10	0/10	5/10	10/10	10/10	10/10	10/10	10/10

Table 18: NYUv2 three-task comparison using a convolutional encoder ($d = 128$, ten seeds, final epoch). Top: single-task references and the three-task model without auxiliaries. Middle: variance-shrinkage auxiliary (*proj*, $\alpha = 1$). Bottom: input reconstruction (*ae*, $\alpha = 250$). Depth RMSE is in meters and normal error in radians; lower is better for both. The recorded "scatter" is $\mathbb{E}_b[\mathrm{Var}_j(h_j^{(b)})]$, with sample correction $d - 1$ and no centering across examples. It differs from (8) and does not establish variation across examples. All rows share the backbone, split, and seeds; absolute values are not directly comparable with published models using different architectures.

Configuration	seg pxacc	seg mIoU	depth RMSE	normal angle	scatter
single-seg	0.359	0.083	2.780	1.574	2.22
single-depth	0.076	0.032	1.159	1.573	5.08
single-normal	0.074	0.031	3.346	0.593	12.43
joint3 (no aux)	0.356	0.100	1.130	0.586	7.30
+ proj aux (α=1)	0.365	0.090	1.085	0.580	1.64
STF-0	0.356	0.100	1.131	0.585	7.30
STFCAL	0.357	0.100	1.134	0.586	7.37
STFCALRANK	0.357	0.095	1.115	0.583	2.65
+ ae aux (α=250)	0.350	0.092	1.168	0.580	3.99
STF-0	0.357	0.099	1.133	0.586	7.19
STFCAL	0.338	0.091	1.171	0.594	3.84

H SUPPLEMENTARY EXPERIMENTAL RESULTS

Ten-seed calibrated experiments support Section 5. At high doses, joint versus STFCAL relative scatter is 0.016 versus 0.990 for HAR ($\alpha = 100$), 0.147 versus 1.001 for battery ($\alpha = 100$), and 0.014 versus 0.982 for RadioML ($\alpha = 3000$). In each case, collapse rates are $10/10$ versus $0/10$. At low doses with closed gates, outcomes are close: HAR at $\alpha = 0.001$ gives 0.525 versus 0.528, and battery at $\alpha = 0.01$ gives 0.963 versus 0.965.

With Transformer encoders at $\alpha = 100$, HAR scatter changes from 0.030 to 1.041 and battery scatter from 0.036 to 1.000, again reducing collapse from $10/10$ to $0/10$. Battery SOH RMSE improves from 0.447 to 0.267. On CelebA at $\alpha = 100$ (five seeds), scatter increases from 0.003 to 0.934 and collapse falls from $5/5$ to $0/5$.

The real-head experiments use $m = 64$ and ten seeds (Table 3). Neither reconstruction head collapses at its tested dose; filtering has little effect or modestly reduces primary accuracy. The training-

Table 19: Intervention families and scope. None of the tested neural-network implementations has a universal trajectory guarantee from the projected norm ratio alone. Historical performance comparisons appear in Tables 1 and 22.

Family	Intervention	Relevant limitation
Task weighting (Chen et al., 2018; Liu et al., 2023)	scalar auxiliary strength	positive attenuation can help at a fixed dose
Gradient geometry (Yu et al., 2020; Liu et al., 2021a; Navon et al., 2022)	task-gradient combination	alignment does not determine variance change
Variance regularization (Bardes et al., 2022)	representation penalty	penalty weight and primary-task trade-off
Full detachment	remove auxiliary encoder gradient	also removes useful auxiliary learning
Selective filtering	remove selected directions	basis identification and subsequent drift
Parameter variance constraint	constrain the actual parameter update	fixed guard batch; finite-step feasibility must be checked

Table 20: Controlled high-dose collapse and task-recovery rates for three domains. Recovery is the fraction of seeds whose primary metric improves over the mean joint-training metric. HAR can retain high accuracy despite substantial shrinkage, so scatter recovery does not imply better accuracy. Battery uses 20 seeds per mode and RadioML ten; the historical HAR file provides aggregate rates. The joint recovery rate is set to zero by convention. All plotted rates are reproduced directly from the released records.

Domain	Metric	JOINT coll. / rec.	STF-0 coll. / rec.	STF-HARD coll. / rec.	STF-SOFT coll. / rec.
HAR	accuracy (flat)	1.0 / 0.0	0.0 / 0.1	0.0 / 0.2	0.0 / 0.1
RadioML	accuracy	1.0 / 0.0	0.0 / 1.0	0.0 / 1.0	0.0 / 1.0
Battery	log-SOH RMSE	1.0 / 0.0	0.0 / 1.0	0.0 / 1.0	0.0 / 1.0

subject classifier presents a different trade-off: joint training achieves identity accuracy 0.646 against chance $1/21$, while filtering reduces it to 0.055 and improves primary accuracy. This is a training-head measurement, not a held-out privacy attack. For high-dose SNR regression, joint training collapses in every seed; hard and soft filtering do not restore accuracy, and detachment performs best among the tested interventions. Probe magnitudes differ substantially across these heads, but the false negatives discussed in Section 4.5 preclude a universal decision rule. Figure 5 shows the battery repair outcome, Table 20 the corresponding collapse and recovery rates, and Table 19 the intervention families.

These experiments delimit the diagnostic and filters. Task metrics may improve despite shrinkage, real heads may be benign at the tested dose, and calibration can produce false negatives. Frozen bases may drift, but the subspace-error bound does not establish drift as the cause of an observed failure. Repeated frozen probes quantify sampling variation rather than certify later training states. Numerical instability and loss of plasticity (Dohare et al., 2024) are additional mechanisms outside the fixed-operator analysis.

The battery encoder maps a per-cycle window $x \in \mathbb{R}^{T \times D}$ of normalized voltage, current, temperature, and resistance to $h \in \mathbb{R}^d$. It combines a TSMIXER feature/temporal-mixing block (Ekambaram et al., 2023) with a gated state-space block (Gu et al., 2022; Gu & Dao, 2024). SOH (L_2), RUL (L_1), reconstruction, and distributional heads share this representation. The real auxiliary heads are evaluated empirically and are not assumed to satisfy Definition 2.

RUL targets require careful handling of right censoring. If a cell has not reached end of life when recording stops, its remaining lifetime is known only to exceed a lower bound. Treating that bound as an exact target can label a healthy cell with zero remaining life. In MATR, 59.2% of samples

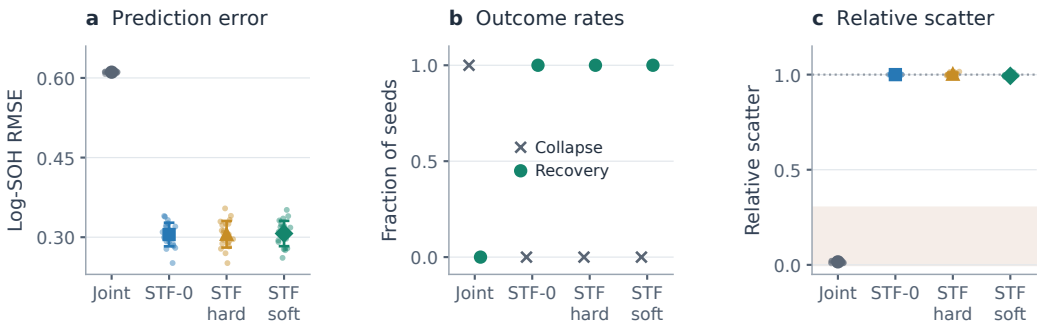

Figure 5: Battery repair under the controlled unified protocol (NASA PCoE, $m = 128$, $\alpha = 10^4$, 20 seeds per mode). (a) Log-SOH RMSE on a linear axis: the mean decreases from 0.611 to approximately 0.305. (b) Collapse and task-recovery rates. (c) Scatter relative to the same-seed single-task reference; the dotted line marks 1.0 and shading marks the operational collapse region below 0.3. Small points are individual seeds; large markers and whiskers show means and population s.d. These log-SOH errors are not directly comparable with percentage-point SOH errors from the separate battery application.

are censored and $96/185$ cells do not reach EOL within their records. Initialization zeros in summary arrays cause a further $48/185$ cells to receive an EOL index of zero under a naive threshold search. Restricting the search to each trajectory's physically valid suffix recovers these cells. We retain censored cells and use $|\hat{r} - r|$ for observed lifetimes and $\max(0, r_{\text{lb}} - \hat{r})$ for censored lower bounds. This one-sided loss is motivated by censored regression (Tobin, 1958), but is not a Tobit likelihood. Excluding cells without observed EOL would leave only 14 test cells in the historical split (Section 5).

To emphasize fast-charge, extreme-temperature, and rapid-degradation windows, BatteryMTL uses three multiplicative weighting factors:

$$\begin{aligned} w_i &= w_i^{(\text{rate})} \cdot w_i^{(\text{temp})} \cdot w_i^{(\text{degrad})}, \\ w_i^{(\text{rate})} &= 1 + \omega\, \mathbf{1}[\text{C-rate}_i > r_q], \\ w_i^{(\text{temp})} &= 1 + \omega\, \mathbf{1}[T_i < t_{lo} \vee T_i > t_{hi}], \\ w_i^{(\text{degrad})} &= 1 + \omega\, \frac{|\Delta\text{SOH}_i|}{\max_j |\Delta\text{SOH}_j|}, \end{aligned} \tag{23}$$

Equation (23) amplifies windows in which several factors are elevated, while also up-weighting windows with a single elevated factor. Thresholds are estimated from the training fold: r_q is the 70th C-rate percentile, and t_{lo}, t_{hi} are the 10th and 90th temperature percentiles. These relative thresholds distinguish regimes within MATR, where 99.8% of cycles exceed 1.5C and temperatures lie within 31–36°C. In contrast, fixed cuts at 1.5C and $0/40^\circ$C would mark nearly all cycles for rate and none for temperature. Section 5 reports regime-specific ablations; SOH-only and RUL-only models provide task-specific references, not guaranteed performance ceilings.

H.1 Composition and Regime Weighting

The following measurements expand the composition and weighting results in Sections 5 and 5.

At the tested high doses, applying PCGrad or CAGrad to the unfiltered battery branch leaves collapse rate at 1.0 and relative scatter at $0.029/0.030$ (Table 21). Hard filtering alone reduces collapse to 0.0, restores relative scatter to 1.00, and improves log-SOH RMSE from 0.611 to 0.294. Adding PCGrad or CAGrad afterward gives RMSE $0.283/0.291$ and scatter up to 1.18. HAR and RadioML show the same qualitative preservation of scatter under composition. These results establish compatibility for the tested order; they neither show that sequencing is necessary nor establish why the unfiltered combiners fail.

Table 21: Filter-then-combiner experiments at high auxiliary doses (five seeds, three domains). Unfiltered branches retain collapse rates of 1.0 and relative scatter at most 0.15. The hard filter, alone or followed by PCGrad/CAGrad, avoids the collapse criterion in every seed. Primary metrics change modestly after composition; the reverse order was not tested. Bold identifies the filter-only reference within each block.

		unfiltered (JOINT)			filtered (STF-HARD)		
	combiner	sum	PCGrad	CAGrad	sum	PCGrad	CAGrad
Battery	log-SOH RMSE↓	0.611	0.592	0.603	**0.294**	0.283	0.291
	collapse rate	1.0	1.0	1.0	0.0	0.0	0.0
	relative scatter	0.02	0.03	0.03	**1.00**	1.05	1.18
HAR	accuracy↑	0.941	0.667	0.952	0.943	0.933	0.940
	collapse rate	1.0	1.0	1.0	0.0	0.0	0.0
	relative scatter	0.15	0.00	0.03	**0.97**	1.14	1.03
RadioML	accuracy↑	0.293	0.091	0.109	**0.301**	0.296	0.295
	collapse rate	1.0	1.0	1.0	0.0	0.0	0.0
	relative scatter	0.02	0.00	0.00	**0.97**	0.99	1.04

BatteryMTL's multiplicative regime weights use training-data percentiles for C-rate and temperature, together with degradation rate. The aggregate SOH results change little, while factor ablations associate improvements in fast-charge and extreme-temperature windows with the rate and temperature terms, respectively (errors 0.95 and 2.35 when the corresponding terms are removed). Table 7 reports the historical aggregate results.

H.2 Probe Dynamics During Training

The frozen calibration in Section 5 is motivated by the behavior of a continuously recomputed probe. We repeat unfiltered joint training with `--rho-trace`, using five seeds per configuration over 1730 HAR steps and 350 battery steps. Monitoring does not alter the measured trajectories: all 28 shared BSCALE configurations match exactly in task metric, collapse rate, relative scatter, and $\hat{\rho}$ (maximum difference zero).

Three observations matter. First, HAR probe values rise sharply during training: from 0.017 to 1.9×10^5 at $m = 32$, 0.010 to 1.0×10^5 at $m = 64$, and 0.006 to 3.5×10^4 at $m = 256$. Second, similar terminal values occur when the toxifier is inactive ($\alpha = 0$) and at the lowest positive dose (10^{-4}), so a large probe value alone does not establish auxiliary-induced contraction. At $m = 32$, increasing the dose instead lowers the terminal probe, to 5.3×10^2 at 10^{-2} and 1.1×10^2 at 3×10^{-2}. Third, in ten HAR configurations whose final scatter lies between 0.50 and 1.04 of the reference, the gate $\alpha\hat{\rho} \geq 1$ opens in every seed between steps 110 and 590. A gate using this running statistic therefore activates on runs that do not meet the collapse criterion at the recorded endpoint.

Battery behaves differently: probe values remain $O(1)$, ranging from 0.29–0.49 initially to 0.3–2.9 at the final step. Among 35 seeds whose final scatter lies above the half-height level, the gate opens only once, at the last step. These domain-dependent trajectories motivate a specified calibration time; they do not establish that a fixed probe is universally preferable to every adaptive rule. The script `repro/scripts/rho_trace_analyze.py` reproduces both analyses.

H.3 Sensitivity to Calibration Time

The default calibration measures $\hat{\rho}$ after four detached warm-up epochs and fixes the gate and rank. We vary only this measurement time: $W \in \{1, 2, 4, 8, 16, 32\}$ at five widths 32–512 for HAR and battery, and $W \in \{1, 4, 16, 32\}$ at four widths for NYUv2, with ten seeds per cell. The $W = 4$ configuration exactly reproduces all 100 shared HAR and battery probe values. NYUv2 gives $2.258/2.008/1.637/1.285$, matching the rounded values in Appendix G.

Table 22: Anti-collapse regularizers at the doses and seeds of Table 1 (ten seeds per configuration). Penalties of weight λ are added to joint training. Entries report collapse rate, primary metric in parentheses, and relative scatter (1.0 is the single-task reference). Joint training, detachment, and hard filtering are evaluated in the same runs. Avoiding the scatter threshold need not improve the primary task.

	HAR		Battery		RadioML	
	coll. / acc.↑	rel. scat.	coll. / RMSE↓	rel. scat.	coll. / acc.↑	rel. scat.
JOINT	1.00 (0.945)	0.15	1.00 (0.610)	0.022	1.00 (0.256)	0.008
STF-0	0.00 (0.945)	1.00	0.00 (0.297)	1.00	0.00 (0.341)	1.00
STF-HARD	0.00 (0.946)	0.99	0.00 (0.299)	1.00	0.00 (0.339)	0.97
VICReg (variance + covariance)						
λ=1	1.00 (0.950)	0.22	1.00 (0.610)	0.023	1.00 (0.249)	0.007
λ=3	0.20 (0.946)	0.31	1.00 (0.610)	0.023	1.00 (0.253)	0.007
λ=10	0.00 (0.944)	0.36	1.00 (0.609)	0.024	1.00 (0.256)	0.008
λ=30	0.00 (0.926)	0.36	1.00 (0.604)	0.033	1.00 (0.260)	0.008
λ=100	0.00 (0.902)	0.36	1.00 (0.451)	0.14	1.00 (0.301)	0.011
λ=300	0.00 (0.873)	0.36	0.00 (0.339)	0.59	1.00 (0.362)	0.035
λ=10^3	0.00 (0.843)	0.36	0.00 (0.390)	1.95	1.00 (0.320)	0.15
λ=3×10^3	0.00 (0.831)	0.36	0.00 (0.423)	3.70	0.10 (0.207)	0.45
λ=10^4	0.00 (0.821)	0.36	0.00 (0.423)	4.54	0.00 (0.215)	1.29
Barlow Twins (redundancy reduction)						
λ=1	1.00 (0.935)	0.11	1.00 (0.607)	0.11	1.00 (0.367)	0.016
λ=3	1.00 (0.765)	0.022	1.00 (0.607)	0.13	1.00 (0.330)	0.024
λ=10	1.00 (0.435)	0.006	1.00 (0.606)	0.15	1.00 (0.121)	0.074
λ=30	1.00 (0.352)	0.006	1.00 (0.606)	0.16	1.00 (0.105)	0.096
λ=100	1.00 (0.338)	0.006	1.00 (0.605)	0.16	1.00 (0.126)	0.072
λ=300	1.00 (0.347)	0.006	1.00 (0.606)	0.16	1.00 (0.141)	0.043
λ=10^3	1.00 (0.343)	0.006	1.00 (0.606)	0.16	1.00 (0.140)	0.042
λ=3×10^3	1.00 (0.365)	0.006	1.00 (0.607)	0.16	1.00 (0.148)	0.053
λ=10^4	1.00 (0.339)	0.006	1.00 (0.607)	0.16	1.00 (0.153)	0.044

Probe values depend on the training state. Across the warm-up sweep, HAR at $m = 32$ increases from 0.63 to 6.9×10^4, and at $m = 512$ from 51.9 to 1.6×10^4. Battery remains $O(1)$: 0.96 to 1.92 at $m = 32$ and 0.058 to 0.67 at $m = 512$. NYUv2 at $m = 32$ changes from 1.94 to 4.19, a factor of 2.2. Accordingly, width slopes fitted to $1/\hat{\rho}$ vary from -2.00 to $+0.55$ for HAR, $+0.37$ to $+0.97$ for battery, and $+0.11$ to $+0.27$ for NYUv2. These are calibration-dependent probe slopes, not interchangeable estimates of capacity scaling. Differences in primary-gradient magnitude, which appears in the denominator, can contribute to this sensitivity.

Measurement time also changes gate activation. For HAR at $m = 64$, the inverse-mean probe value is 1.34 at $W = 1$, 0.181 at $W = 2$, 9.07×10^{-3} at $W = 4$, and 1.7×10^{-5} at $W = 32$, compared with the measured transition interval $[0.001, 0.01]$. In the sixty-epoch, ten-seed experiment at $\alpha = 0.03$, gates remain closed for $W = 1$ and $W = 2$: both settings collapse in $10/10$ seeds, with accuracy 0.947 and relative scatter $0.19/0.20$, close to joint training (0.949, 0.19). For $W \geq 4$, all gates open, no seed collapses, and scatter returns to 0.99–1.00. At $\alpha = 100$, all six warm-up settings activate the gate and avoid collapse.

A short warm-up therefore fails at the marginal dose in this experiment. Detached warm-up also changes the training trajectory, so closed-gate runs need not equal joint training from initialization. Checking whether $\alpha\hat{\rho} \geq 1$ verifies activation at the chosen operating dose; it does not independently validate protection. The script `repro/scripts/warmup_sensitivity.py` reproduces these results.

I ANTI-COLLAPSE REGULARIZERS AT MATCHED DOSES

Table 22 evaluates two regularizer families at the doses and seeds of Table 1: VICReg's variance hinge plus off-diagonal covariance penalty, and Barlow Twins' redundancy-reduction penalty. Each is applied to the shared representation with weight λ. VICReg's invariance term is omitted because this experiment has no augmented-view pair. All rows use the same joint-training baseline; these are adaptations of the regularizers to this setting, not evaluations of their complete original learning pipelines.

Regularization can clear the operational collapse threshold without restoring the single-task reference or improving the task. In the coarse grid, VICReg first avoids collapse in all seeds at $\lambda = 10$ for HAR, 300 for battery, and 10000 for RadioML. Barlow Twins does not do so at any tested weight.

On HAR, VICReg scatter stays near 0.36 from $\lambda = 10$ to 10^4, while accuracy falls from 0.944 to 0.821. At the lowest clearing weight, accuracy is close to joint training (0.945) and detachment. On battery, scatter increases through the reference level, but the best tested RMSE is 0.339 at $\lambda = 300$ (relative scatter 0.59), compared with 0.297 for detachment. Larger weights worsen RMSE to 0.390 at 10^3 and 0.423 at 10^4; detachment is better in nine of ten seeds. On RadioML, $\lambda = 10000$ gives zero collapse and scatter 1.29, but accuracy drops to 0.215, below both collapsed joint training (0.256) and detachment (0.341). These cases separate the scatter criterion from predictive utility.

The Barlow Twins adaptation yields absolute scatter as low as 0.011, 0.012, and 0.013 for HAR, battery, and RadioML, respectively, or factors of 172, 6, and 24 below their references. It fails to clear the collapse threshold throughout the tested grid, with HAR accuracy reaching approximately 0.34–0.44 and RadioML approximately 0.11 in the low-accuracy configurations.

Additional ten-seed sweeps use $\lambda \in \{2, 5, 20, 50, 200, 500\}$ on HAR and $\lambda \in \{200, 500, 700, 1500, 2000, 5000\}$ on battery. The recorded battery transition is bracketed by 150 and 200, and the lowest measured battery RMSE remains 0.339 at $\lambda = 300$, above detachment. HAR scatter remains near 0.36 above its crossing. RadioML was not refined: the sampled collapse rate is 1.0 at 1000, 0.10 at 3000, and 0.0 at 10000. The experiments therefore locate an operational transition over a finite grid rather than a precise critical weight.

The collapse criterion thresholds relative scatter, whereas the regularized solution depends on the data, model, penalty, and its weight. Nothing in these objectives forces the resulting scatter to match the single-task reference. The tested grid therefore supports reporting representation statistics and task error together, rather than treating either as a sufficient measure of recovery.