WHAT MUST SURVIVE A SUBSPACE REFRESH? OPTIMIZER MEMORY, CALIBRATION, AND INSTABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Refresh-based low-rank optimizers periodically change the low-dimensional representation in which they maintain gradient statistics. When this representation changes, must the accumulated statistics be preserved, or can they be rebuilt from new gradients? We vary the statistical age used in Adam's standard second-moment bias correction. At each refresh, we erase the second-moment statistic v while retaining the first moment m, which smooths gradients. We compare two reconstructions that discard the same pre-refresh history and differ only in whether the correction age is global or restarted. The global-age reconstruction raises final perplexity to roughly $64\times$ the no-reset baseline on average across three GaLore-1B runs. Using the rebuilt statistic's own age recovers carry-level quality across GaLore scales up to 1B parameters, and in PLUMAGE and Fira runs at 60M, despite discarding all pre-refresh v history at every refresh. The failure of the global-age ablation therefore cannot be attributed to history loss alone: preserving pre-refresh v history is not necessary to recover carry-level quality in these settings. We also show that calibrated reconstruction can become hazardous from some training states, and trace one mechanism by which this occurs. In GaLore-60M, repeated oversized updates can drive attention concentration and highly intermittent gradients, causing a freshly rebuilt second moment to misrepresent the scale of retained momentum. Targeted interventions support this feedback mechanism. A late switch to correctly aged reconstruction progressively improves the gradient regime and training quality. A correct statistical age is therefore sufficient along the healthy trajectories tested, but calibration alone does not make reconstruction state-invariant: safe reconstruction also depends on the state training has reached.

1 INTRODUCTION

Adam (Kingma & Ba, 2015) determines each update from the current gradient and statistics accumulated over previous gradients. This history helps smooth update directions and set their magnitudes, but storing it for all parameters accounts for a substantial share of training memory. That history is not incidental: because the second moment sets the scale of every step, misjudging it yields steps far too large or too small. Low-rank optimizers reduce this cost by representing gradients and the associated optimizer state in a lower-dimensional subspace. Some of these methods (Zhao et al., 2024; Robert et al., 2025; Chen et al., 2025; Haroush & Soudry, 2025; Hao et al., 2024; Xiao et al., 2025; Rajabi et al., 2025b) periodically refresh the subspace to keep its directions aligned with the evolving gradients. Each refresh therefore raises a basic question: when the representation used for optimization changes, what information from the optimizer's past must actually survive? These methods must decide what to do with the accumulated statistics when the coordinates change—carry them, transport them, realign them—and these design choices are commonly justified by removing state and observing whether training degrades. Such a removal, however, changes two things at once: it discards the history of a statistic and, unless the rebuilt statistic is recalibrated, changes its effective scale. The ablation is therefore only diagnostic of historical necessity if the replacement is otherwise comparable. We exploit this to vary the two separately, and show that a correct statistical age is sufficient to recover carry-level quality, but that calibration alone does not make reconstruction safe to apply from every training state.

Adam maintains two statistics: the first moment m, which smooths gradients, and the second moment v, which helps control update magnitude. To test the role of this history, we compare several reset strategies across four optimizers: GaLore (Zhao et al., 2024), LDAdam (Robert et al., 2025),

Initial code release on GitHub: see Appendix A.

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.

PLUMAGE (Haroush & Soudry, 2025), and Fira (Chen et al., 2025). Resetting only the second moment v at every refresh severely degrades training. Across three GaLore-1B runs, final perplexity reaches roughly $64\times$ the matched no-reset baseline on average. Yet resetting the first moment m as well substantially reduces this damage. The amount of discarded information alone therefore cannot explain the outcomes. This asymmetry motivates our focus on the most vulnerable case: rebuilding v while retaining m, which keeps the numerator history fixed while isolating the fate of the second-moment history.

Adam's second moment v is an exponential moving average of squared gradients. Starting from zero biases this average downward, so Adam divides it by a correction factor that depends on the number of updates since initialization. That a statistic rebuilt from zero requires its own correction is exactly what this mechanism was designed for; we do not treat it as a finding, but as a known and exactly predictable lever. A v rebuilt at a refresh has accumulated only 0.1% of its steady-state mass after one update; correcting it with the global counter treats it as mature, and because the update is divided by its square root, the first post-refresh update is amplified by up to $\sim 32\times$. A conventional reset thus changes two things at once: it removes the pre-refresh history of v but leaves the second moment's bias-correction counter unchanged, treating the newly rebuilt statistic as if it had accumulated gradients since the start of training. This amplification follows in closed form from the two ages (Eq. 1) and is matched within 0.1% in our runs, so the miscalibrated reset acts as a controlled perturbation. To separate the two effects, we remove exactly the same v history but give the rebuilt statistic its own statistical age. This change alone recovers the quality achieved by retaining state in complete GaLore runs on LLaMA-60M, 130M and 1B (Touvron et al., 2023), despite discarding all pre-refresh v history at every refresh. In the GaLore settings we test, the complete pre-refresh history of v is therefore not necessary to recover carry-level quality. What a reset ablation establishes thus depends on how the replacement is calibrated: removing a statistic and observing failure does not show that its history was necessary if the ablation also miscalibrates its replacement.

Is calibrated reconstruction then safe from every state training can reach? The miscalibrated reset gives us a way to ask. Across 192 GaLore-60M runs spanning 64 (β_2, T) settings, outcomes split into two populations—near baseline or severely degraded—with few cases in between, and near the boundary identical hyperparameters and data can land in either. From one such degraded state, the same calibrated reconstruction strongly amplifies updates. Since projectors are recomputed from the gradients, one might expect the damage to propagate primarily through subspace selection. Yet this is not the whole story: the SVD never sees v, and the acute damage largely disappears when we reduce update magnitude without changing its direction. Instead, repeated oversized updates drive a small set of attention heads toward near-deterministic concentration; such heads produce gradients only in occasional bursts, so gradients become highly intermittent. A freshly rebuilt v can then reflect an unusually weak current gradient even while retained m still carries recent bursts, producing another oversized update. Frozen-state, sustained, amplitude-only and projector interventions probe each link of this chain, and a late switch to local-age reconstruction progressively restores the gradient regime and training quality while projectors remain endogenous. The trajectory itself thus becomes part of what later reconstruction must contend with.

2 Related work

Existing methods address these challenges through state transfer or subspace evolution. GaLore (Zhao et al., 2024) directly reuses the projected moments; its variants modify the projection rule, numerical precision, or refresh frequency (Hao et al., 2024; Zhang et al., 2025; Das, 2024; Chen et al., 2025; Su et al., 2025). LDAdam (Robert et al., 2025) transports representable state into the new basis and compensates for compression loss through error feedback; PLUMAGE (Haroush & Soudry, 2025) realigns the state; COAP (Xiao et al., 2025) promotes overlap between successive subspaces. Other methods act on subspace selection through random projections (He et al., 2025) or Grassmannian tracking (Rajabi et al., 2025b;a). These approaches seek to keep optimizer state usable as the representation changes. Thomas (2026) studies how to adapt Adam's statistics when subspace directions change. This work shows that the information in v is generally insufficient to recompute the second moment exactly in the new coordinates. It also finds that giving less weight to older gradients can sometimes improve training. Our question is complementary: can this history be discarded entirely and v reconstructed without losing the quality achieved by retaining state? SPAM (Huang et al., 2025) periodically resets both Adam moments and uses a short post-reset warmup, while SMET (Xiao et al., 2026) acts on optimizer state during parameter regrowth. Failures after optimizer-state resets have also been reported and addressed through the learning-rate schedule (Lialin et al., 2024; Huang et al., 2025). Unlike such joint-reset/warmup schemes, we

retain m and the original learning-rate schedule, isolating second-moment history removal from statistical-age calibration before tracing how miscalibrated reconstruction reshapes the gradient regime and future representations.

3 EXPERIMENTAL SETUP

We first compare interventions on Adam's state, then examine the role of subspaces and their refresh frequency. We study GaLore, LDAdam, PLUMAGE, and Fira, with multi-scale experiments on GaLore using LLaMA-60M, LLaMA-130M, and LLaMA-1B, trained on C4/en with the t5-base tokenizer. Matched full training runs share initialization, data order, and the learning-rate schedule, without retuning across variants. One-step interventions start from the same state. Final quality is evaluated by perplexity on the C4/en validation split. Appendix A provides configurations, detailed protocols, and the exact intervention placement for each method. LDAdam requires two distinct experiments because it refreshes its subspace at every update: the asymmetry study applies resets at every refresh, whereas the local-age study uses a single reset at update 2,000 to allow v to rebuild afterward. For `reset_m` and `reset_all`, the raw first moment is zeroed but its bias-correction age remains global. The two LDAdam protocols are analyzed separately.

Table 1: Main state interventions at a refresh. *Retained* means that the state follows the method's native transition rule, not that it remains unchanged.

	baseline	`reset_m`	`reset_v`	`reset_all`	`local_v_age`
m	retained	$\leftarrow 0$	retained	$\leftarrow 0$	retained
v	retained	retained	$\leftarrow 0$	$\leftarrow 0$	$\leftarrow 0$
age of v	global	global	**global**	global	**local**

4 SEPARATING HISTORY LOSS FROM RECONSTRUCTION MISCALIBRATION

We use `reset_v` as a diagnostic ablation of second-moment history, not as a proposed training rule: its failure illustrates that removing optimizer state can confound history loss with miscalibration of the replacement state. With GaLore on LLaMA-130M, resetting the second moment v at every refresh raises final perplexity to 17.51 times the matched no-reset baseline. Resetting the first moment m as well brings this ratio down to $1.40\times$. Training is therefore less affected by removing both moments than by removing v alone. This ordering holds across all four tested optimizers, despite their different state-transition rules. Degradation after resetting v also appears at a larger scale: across three LLaMA-1B GaLore runs, `reset_v` reaches on average $64\times$ the perplexity of its matched baseline ($68.5\times$, $60.3\times$ and $63.2\times$ for seeds 0–2). Resetting the first moment m alone leaves final perplexity within roughly 1% of baseline in the five configurations where it is tested. If losing v history were the whole explanation, discarding m as well should not reduce the damage so sharply. We therefore focus on the most vulnerable regime, rebuilding v while retaining m: this keeps the numerator history fixed and isolates how the rebuilt denominator is calibrated. In the conventional reset, `reset_v`, Adam uses the total number of training steps to correct the rebuilt statistic. As a calibrated control, `local_v_age` discards the same pre-refresh v history at the same steps but resets the correction counter to count updates since the reset. The bias correction for m and the placement of interventions within the transition remain identical.

Table 2: Final-perplexity ratios to matched no-reset baselines, with resets at every refresh. Ratios use mean perplexities over three realizations per condition (two for Fira). Details: Appendix B.

	GaLore 60M	GaLore 130M	LDAdam 60M	PLUMAGE 60M	Fira 60M
`reset_m`	1.00	1.00	1.00	1.00	1.01
`reset_all`	1.36	1.40	4.41	1.39	4.57
`reset_v`	11.82	17.51	34.99	9.54	22.40

Repeated local-age reconstruction recovers carry-level quality despite discarding all pre-refresh v history at every refresh: within 1.5% of carry at 60M/130M and 0.6% at 1B across three matched realizations per scale (Table 3 and Table 6), and at carry level for PLUMAGE and Fira at 60M (Appendix B). Preserving pre-refresh v history is therefore not necessary to recover carry-level quality in the tested settings. The local-age endpoints lie slightly below carry; the shared schedule is not retuned, and we do not interpret this gap. Nor does this result imply that the same reconstruction is benign from every state training can reach; Sections 6 and 7 take up that question. Recovering

Table 3: GaLore-1B final validation perplexity, three matched realizations, shared 100k-step schedule (not retuned).

Run	Carry	Local-age
Seed 0	14.8766	14.8188
Seed 1	14.8643	14.7847
Seed 2	14.8171	14.7823

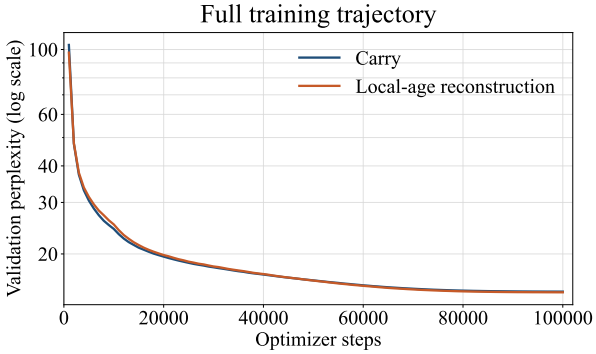

Figure 1: LLaMA-1B, seed 1: local-age reconstruction tracks carry despite discarding v at every refresh.

baseline quality does not mean recovering the same state. After each reset, the first moment m retains the influence of older gradients that the second moment v no longer records. Figure 1 shows one representative uninterrupted matched LLaMA-1B run.

5 Statistical age and optimizer-state reconstruction

We first analyze reset-induced amplification within a fixed representation, including the effects of resetting m and retaining part of v. We then separate statistical calibration from history compatibility, asking which properties of the discarded state are actually required to recover stable training. Finally, we note that each refresh is selected from a trajectory shaped by earlier updates, a feedback that Section 7 tests in the full optimizers.

5.1 Rebuilding v under a mature clock

When v is erased and rebuilt, two clocks disagree: the statistic is one update old, the optimizer counter is not. The immediate scaling effect follows from that gap. Let s be the optimizer age immediately before refresh and $t \geq 1$ the number of updates since. At $(\beta_1, \beta_2) = (0.9, 0.999)$ the reconstruction timescales of the two moments are $\tau_m = 10$ and $\tau_v = 1000$ updates. Rebuilt from zero, the second moment has accumulated $v_t^{\text{rec}} = (1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_i^2$ after t updates. At the first step, its accumulated EMA mass is only $1-\beta_2 = 10^{-3}$. That same raw statistic can be corrected by the global age, $\hat{v}_t^{\text{global}} = v_t^{\text{rec}}/(1-\beta_2^{s+t})$, or by its own, $\hat{v}_t^{\text{local}} = v_t^{\text{rec}}/(1-\beta_2^t)$. For identical raw moments and $\epsilon = 0$, changing only the second-moment correction age gives the exact relation

$$\Delta_t^{\text{global}} = A_{\text{eff}}(t; s)\, \Delta_t^{\text{local}}, \qquad A_{\text{eff}}(t; s) = \sqrt{\frac{1-\beta_2^{s+t}}{1-\beta_2^t}}, \tag{1}$$

with no assumption on the gradient distribution. At matched raw states, local-age correction is a post-reset step-size multiplier A_{eff}^{-1} fixed by β_2 and the two ages. This remains exact for nonzero ϵ when second-moment correction enters the scalar step multiplier, as in GaLore. When ϵ is added after correction inside the denominator, the ratio becomes coordinate-dependent. Numerical predictions use each executed rule. Calling the local branch *calibrated* requires one further assumption. If the post-refresh gradients have local second moment $\mathbb{E}[g_i^2] = \sigma'^2$, then $\mathbb{E}[v_t^{\text{rec}}] = (1-\beta_2^t)\sigma'^2$, so $\mathbb{E}[\hat{v}_t^{\text{local}}] = \sigma'^2$ while $\mathbb{E}[\hat{v}_t^{\text{global}}]$ falls short of it by the factor $(1-\beta_2^t)/(1-\beta_2^{s+t})$. A mature clock, under this model, underestimates a freshly rebuilt second moment. In the mature limit the first global-age adaptive update is amplified by $31.6\times$ relative to the local-age branch; the factor stays above $5\times$ for about 41 updates and above $2\times$ for about 288, so at the reference interval $T = 200$ it has not finished decaying when the next refresh arrives. Across four matched healthy refreshes, measured global/local update-norm ratios agree with predictions within 0.1% (Appendix C.1). This agreement verifies the algebra and its implementation. Does correcting the mismatch also remove the immediate shock? At matched states, local-age correction reduces the median reset update from $25.32\times$ to $1.23\times$ the matched baseline and removes the median same-minibatch loss increase.

The same mechanism appears under a different transition rule. LDAdam refreshes at every update, so resetting v at every refresh would leave it no history to accumulate. We therefore apply a single reset at update 2,000 and compare two runs differing only in the age used for v. In an instrumented replication the update under the global age is 3.12 times larger, matching the calculation that retains ϵ; training loss overshoots its pre-reset level by 0.177 and returns to its pre-reset level after 39 updates,

whereas under the local age no overshoot appears in the raw training loss. Both ages yield similar final perplexities (Table 8). This protocol therefore identifies the local mechanism under a different transition rule, not repeated reconstruction as an alternative to carry.

5.2 Why erasing m as well softens the shock

Resetting both moments also shrinks the numerator, initially offsetting the underestimated denominator. Because m rebuilds about 100 times faster than v, this attenuation is short-lived. Assume the post-refresh mean gradient μ' is locally constant and the carried first moment calibrated to it, $\mathbb{E}[m_{\text{pre}}] = (1-\beta_1^s)\mu'$ — an assumption about compatibility with the *new* representation, not a consequence of calibration in the old one. Under it, the retained numerator of the `reset_v` branch stays unbiased after global bias correction. Our reference magnitude U_{ref} uses the ratio of the expected corrected numerator to the square root of the expected corrected second moment, both calibrated. Erase m along with v and its expected corrected numerator is reduced relative to the calibrated numerator by $(1-\beta_1^t)/(1-\beta_1^{s+t})$. Combined with the second-moment factor of equation 1,

$$A_{\text{all}}(t;s) = \frac{1-\beta_1^t}{1-\beta_1^{s+t}} A_{\text{eff}}(t;s), \tag{2}$$

where A_{all} is the mean-field update magnitude relative to U_{ref} and the prefactor is below one for $s > 0$. At the first step, for a mature optimizer, the amplification is $3.16\times$ against $31.6\times$ under `reset_v`. The attenuation comes from the factor $1-\beta_1$ in the rebuilt numerator, and because $\tau_m \ll \tau_v$ it fades much faster than the denominator mismatch. At the matched median refresh the realized norms are $12.1\times$ and $44.8\times$ the matched baseline under joint reset and `reset_v`, so the joint-reset update carries 27% of the norm. The formula correctly predicts the direction of the attenuation. Its magnitude is not expected to match the mean-field ratio, since it depends on the carried numerator, coordinate heterogeneity, and ϵ.

5.3 Partial retention provides a calibration floor

We treat a single refresh from calibrated state here; repeated retention is treated separately in Appendix C.3. Consider one projected coordinate, with second-moment scales σ^2 before and $\sigma'^2 > 0$ after the refresh, and write $\rho = \sigma^2/\sigma'^2$ for their ratio. Let $\gamma \in [0,1]$ be the retained fraction, so $v_0^+ = \gamma v_{\text{pre}}$, and for $\gamma > 0$ assume the retained state calibrated to the old scale, $\mathbb{E}[v_{\text{pre}}] = (1-\beta_2^s)\sigma^2$. For a common ambient gradient distribution, let $S = \mathbb{E}[zz^\top]$, where z is the gradient vector being projected. Treating the pre- and post-refresh directions u, u' as fixed gives $\sigma^2 = u^\top S u$ and $\sigma'^2 = u'^\top S u'$.

Proposition 1 (Second-moment reconstruction)**.**

$$\frac{\mathbb{E}[\hat{v}_t]}{\sigma'^2} = \frac{\gamma\rho(1-\beta_2^s)\beta_2^t + (1-\beta_2^t)}{1-\beta_2^{s+t}}, \qquad A(t;s,\gamma,\rho) = \left(\frac{\mathbb{E}[\hat{v}_t]}{\sigma'^2}\right)^{-1/2}. \tag{3}$$

The numerator adds the surviving contribution of old gradients to the contribution accumulated since the refresh. For $\gamma = 0$ the inherited term vanishes and equation 1 follows, independently of the old scale — no assumption on the retained state is needed, so the formula holds even when earlier resets have left v_{pre} uncalibrated. For $\gamma > 0$, in the mature-age limit $s \to \infty$, the normalized second moment becomes $\gamma\rho\beta_2^t + (1-\beta_2^t)$, bounded below by $\min(\gamma\rho, 1)$ whenever $\gamma\rho > 0$. Keeping a positive inherited contribution therefore puts a floor under the mean-field second moment. When the historical term dominates, at fixed numerator and negligible ϵ, the update ratio relative to the $\gamma = 1$ carry branch is approximately $\gamma^{-1/2}$: 1.41 and 2.00 for $\gamma = 0.5$ and 0.25, measured at 1.40 and 1.96. Among the tested retention levels, every $\gamma \geq 0.25$ ends within $1.04\times$ the matched baseline over complete runs, whereas $\gamma = 0$ enters the catastrophic regime. This reproduces the ordering of Section 4: even partial inherited mass keeps training out of the regime that complete erasure reaches.

5.4 What local age changes, and what it does not

Local-age correction changes normalization without restoring discarded history. To distinguish these effects, we express each moment as a weighted sum of past gradients. For the scalar EMA core considered here — state evolving through EMA updates and the interventions of Section 3, without coordinate mixing — write $m = \sum_i a_i g_i$ and $v = \sum_i b_i g_i^2$ with $a_i, b_i \geq 0$; no stationarity or independence assumption is needed. With $A = \sum_i a_i$, $B = \sum_i b_i$, $p_i = a_i/A$, $q_i = b_i/B$, and c_1, c_2 the bias corrections applied to m and v, the adaptive core at $\epsilon = 0$ factors as

$$\frac{\hat{m}}{\sqrt{\hat{v}}} = \underbrace{\frac{A}{c_1}\sqrt{\frac{c_2}{B}}}_{\text{normalization}} \quad \underbrace{\frac{\sum_i p_i g_i}{\sqrt{\sum_i q_i g_i^2}}}_{\text{normalized histories}}. \tag{4}$$

Proposition 2 (History compatibility)**.** *Assume $A, B > 0$. Then*

$$\sup_{\{g_i\}\,:\,v>0} \frac{\hat{m}^2}{\hat{v}} = \frac{A^2 c_2}{B c_1^2}\,\mathcal{K}(p,q), \qquad \mathcal{K}(p,q) = \begin{cases} \sum_{i:q_i>0} p_i^2/q_i, & \operatorname{supp}(p) \subseteq \operatorname{supp}(q), \\ +\infty, & \textit{otherwise,} \end{cases} \tag{5}$$

where support *refers to historical gradient contributions, not parameter coordinates. Method-specific state transports that mix coordinates fall outside this representation.*

When finite, $\mathcal{K}(p,q) = 1 + \chi^2(p\|q)$, with χ^2 the Pearson divergence between the normalized historical weights; $\mathcal{K} = 1$ exactly when they coincide. The bound is worst-case and structural: it says what a gradient sequence could do, not what training does. A finite bound requires every past gradient contributing to m also to contribute to v (proof in Appendix C.4). This characterizes a structural difference between the reset rules. After the first post-reset update, `reset_m` retains $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$ and `reset_all` rebuilds a common support, whereas `reset_v` violates the inclusion by retaining history only in the numerator. Partial retention behaves like the former: starting from compatible histories, any $\gamma > 0$ keeps the old contributions represented in v and so keeps $\mathcal{K}$ finite, whereas $\gamma = 0$ removes them entirely. Complete erasure is therefore qualitatively distinct from arbitrarily small positive retention — though a finite $\mathcal{K}$ need not be a small one. Before the matched global- and local-age branches are allowed to diverge, A, B, p and q are identical and only c_2 differs. At $\epsilon = 0$ the intervention therefore acts through the normalization factor alone, leaving the normalized-history term untouched; the implementation-specific relation for nonzero ϵ is in Appendix C.2. Both `reset_v` and `local_v_age` have $\mathcal{K} = \infty$: they discard the same history and lose the same finite worst-case bound. Yet only local-age reconstruction recovers carry-level quality in the tested GaLore runs. Changing the correction age therefore restores training quality without restoring history compatibility. The loss of the bound alone cannot explain the different outcomes. Calibration therefore does not make the discarded history irrelevant; it shows that preserving it is unnecessary along the states visited by these runs. Whether this reconstruction is safe to apply from every training state is the question the rest of the paper pursues.

So far, the representation has been held fixed. In practice, however, each refresh is chosen from a trajectory shaped by earlier updates. Reconstruction can therefore change not only the next update, but also the representations on which later state is rebuilt. Appendix C.5 gives a constructive stability example; Section 7 traces this effect in the language-model runs.

6 Where does miscalibrated reconstruction lead?

Because its immediate effect is known, the miscalibrated reset can be used to reach states that calibrated reconstruction never visits on its own; this section maps which states it reaches. Our starting question was whether the pre-refresh history stored in v must survive a subspace refresh. The previous sections show that its apparent necessity can instead arise from miscalibrating the rebuilt statistic. This leaves a second question: when does that miscalibration remain a local perturbation, and when does it develop into persistent training failure? Two parameters follow naturally from the analysis above. β_2 controls both the magnitude of the reset-induced amplification and how quickly it decays, while T sets the number of updates before v is reconstructed again. Their interaction therefore controls how strongly successive reconstruction shocks overlap. Neither is an isolated control, however: β_2 also changes the baseline dynamics of v, and under our protocol T changes subspace-refresh frequency together with reconstruction frequency. We sweep 64 (β_2, T) configurations with GaLore–LLaMA-60M, using three matched realizations per configuration for 192 trajectories (Figure 2). This is not a search for favorable hyperparameters for a reset rule we already know to be miscalibrated; it asks when that known perturbation remains tolerable and when it develops into persistent failure.

The outcomes separate sharply into two regimes: most trajectories either remain close to their matched baseline or deteriorate severely, with relatively few intermediate cases. We define severe degradation as final perplexity exceeding $5\times$ the matched baseline. The reference setting analyzed in Section 7, $(\beta_2, T) = (0.999, 200)$, lies inside this severe region rather than being an isolated failure chosen for its magnitude.

The broad structure of Figure 2 follows the timescales predicted by the analysis. As β_2 increases, reset-induced amplification decays more slowly, while larger T gives it more time to dissipate before the next reconstruction. At $\beta_2 = 0.999$, the analysis still predicts about $2.3\times$ residual amplification immediately before the next reset even at $T = 200$, consistent with no tested T leaving the severe region at that β_2. This account is qualitative rather than a cell-by-cell prediction; $(\beta_2, T) = (0.95, 20)$ degrades more than it suggests.

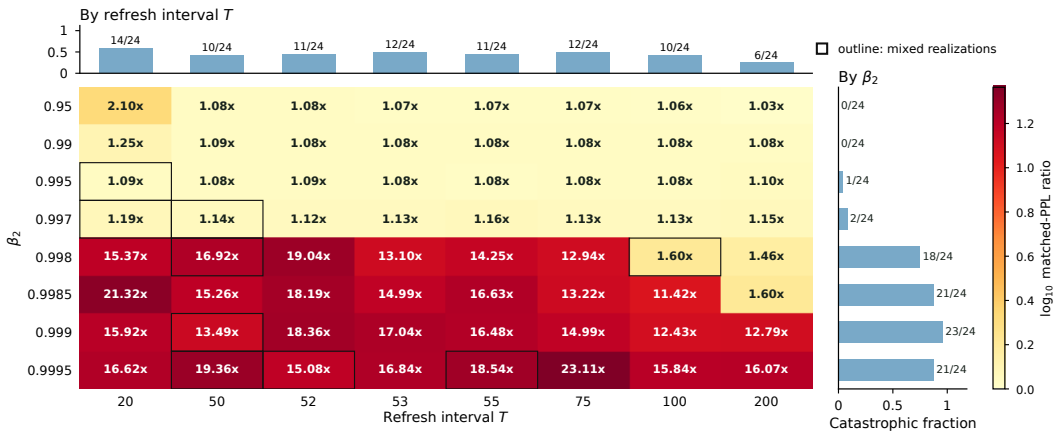

Figure 2: GaLore-60M, global-age `reset_v`: median final-perplexity ratio to matched baseline over three realizations per (β_2, T). Outlined cells mix severe ($> 5\times$) and non-severe outcomes; marginals give the severe fraction. The reference setting of Section 7, $(0.999, 200)$, is severe.

Table 4: Diagnostics in the reference run. *Attn. norm.*: sustained layer-0 Q/K logit normalization from update 2,200, OWN projectors. Healthy carry: PPL ≈ 32 at 4.8k. Full values: Tables 10, 11, 14.

	OWN 2k	OWN 4.8k	REPLAY 4.8k	Attn. norm. 4.8k
Reset sensitivity R	0.25	10.8	0.16	1.4
Mismatch Ψ	1.3	14.8	1.5	2.8
Low-gradient momentum	10%	78%	11%	25%
Gradient concentration	38%	90%	41%	53%
Attention logit scale	2.4	10.2	3.0	4.2
Attention entropy	.43	.07	.39	.27
Validation PPL	48	72	35	39

But β_2 and T do not determine the outcome on their own. Near the boundary, identical (β_2, T) values can produce either a near-baseline or a severely degraded trajectory. Hyperparameters therefore set susceptibility but do not determine which regime a given trajectory reaches.
If reset damage were purely cumulative, one might expect outcomes to vary more gradually with the residual amplification. The observed separation is therefore consistent with a trajectory-dependent feedback process, which Section 7 tests directly.

7 WHY RECONSTRUCTION CAN REMAIN HAZARDOUS AFTER CALIBRATION

Section 6 shows that identical reset conditions can lead to either regime, so factors realized along the trajectory matter. The question now shifts from whole runs to states. Repeated calibrated reconstruction recovers carry-level quality when used throughout training; is the same calibrated reconstruction state-invariant, i.e. equally safe from any state training can reach? Here we identify one route by which it need not. Repeated oversized updates change the gradient regime itself: attention concentrates, gradients become intermittent, and a freshly rebuilt v can then misrepresent the scale of the retained m (Figure 3). Once in this regime, even correctly aged reconstruction can strongly amplify updates. Values below come from the reference GaLore–LLaMA-60M run; the matched OWN/REPLAY transition and the main interventions replicate in direction and approximate magnitude in two additional realizations (Appendix D).

Branches and diagnostics We compare two runs that apply the same uncorrected `reset_v` rule and differ only in where their projectors come from: OWN recomputes them from its own perturbed trajectory, whereas REPLAY uses those recorded from the matched healthy run. Differences between the two therefore reflect projector provenance alone. Each diagnostic answers one question. *Is reconstruction itself hazardous at this state?* The reset sensitivity

$$R_k = \left\| \Delta_k^{\text{local-age reset}} \right\| / \left\| \Delta_k^{\text{no reset}} \right\| \tag{6}$$

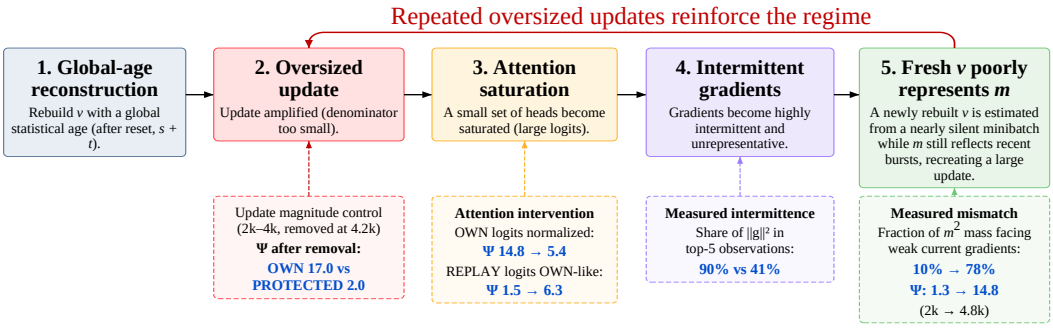

Figure 3: Mechanism chain in the reference GaLore-60M run ($\beta_2 = 0.999, T = 200$). Solid boxes: proposed mechanism; dashed boxes: measurements or interventions probing each link; arrow: oversized updates reinforce the regime.

compares, from the same state, the update after a correctly aged reset with the update without reset: $R < 1$ means reconstruction shrinks the update, $R \gg 1$ that it amplifies it. The reported R uses realized parameter displacements across all 56 GaLore matrices (Appendix D). *Is the current gradient still a good scale for the retained momentum?* The momentum–gradient mismatch

$$\Psi_k = \left\|\Delta_k^{\text{fresh}}\right\| / \left\|\Delta_k^{\text{recent}}\right\| \tag{7}$$

holds m fixed and normalizes it either by the current gradient or by the preceding 200 gradients. $\Psi \approx 1$ means the two scales agree; a large Ψ means the current gradient provides too little scale for the retained momentum, inflating the fresh-denominator update. *What changes in the gradient regime? Low-gradient momentum* is the fraction of m^2 in coordinates whose current gradient magnitude is below $0.1\times$ its recent scale, and *gradient concentration* is the share of squared-gradient energy carried by the 5 largest of 64 recent observations; high values of both indicate intermittent gradients. Attention logit scale and entropy measure how concentrated the selected heads are. Ψ, low-gradient momentum, and gradient concentration use layer-0 Q/K coordinates; attention metrics use the selected heads, while R covers all 56 GaLore matrices (Appendix D.1).

7.1 Attention-logit interventions modulate the intermittent regime

In late OWN, attention in the selected heads has collapsed (logit scale 2.4 to 10.2, entropy 0.43 to 0.07) and gradients arrive in bursts: five of 64 recent observations carry 90% of the squared-gradient energy, against 41% under REPLAY. Is attention a cause? At the frozen update-4,800 state, rescaling only the selected Q/K logits acts in both directions: making OWN logits REPLAY-like lowers Ψ from 14.8 to 5.4, and the reverse raises it for REPLAY from 1.5 to 6.3 (Appendix D.3). Applied throughout training from update 2,200 under the archived OWN projector sequence, the same normalization keeps Ψ at 2.8 at update 4,800, lowers R over all 56 GaLore matrices from 10.8 to 1.4, and improves perplexity from 72 to 39 (Table 4). Intervening on these logits therefore causally modulates the late reconstruction-sensitive regime; because the transformed logits remain in the backward graph, the intervention does not isolate forward attention concentration from the induced change in backward gradients.

7.2 Oversized updates create the regime

Is the damage due to update size or direction? At the late pathological state, shrinking the update to 7% of its original norm using a 32-minibatch denominator almost removes the loss increase, and keeping the original direction at a comparable norm does the same (Appendix D.5): the acute damage is an amplitude effect. Amplitude also matters upstream. Controlling only the update norm from update 2,000 to 4,000, direction unchanged, keeps the regime from forming, and the effect outlasts the control: 1,400 updates after its removal, Ψ is 2.0 instead of 17.0 and logit scale 3.7 instead of 11.1 (Appendix D.6). Oversized updates therefore do more than cause bad steps: they push training into the regime that makes later reconstruction hazardous.

7.3 Future projector sequences affect persistence

Projectors do not create the shock but influence how long it lasts. We branch from the pathological state just before the refresh at update 4,800 and measure the reset-boundary update norm 800 updates later. Swapping in the healthy projector at this single refresh provides limited relief (228 to 176); imposing the healthy projector sequence from that refresh onward reduces it to 21 (Appendix D.7). The two interventions differ in duration, so they do not rank current against future projectors. They

do show that the future representation sequence can causally alter persistence of the pathological regime, supporting the representation-feedback interpretation in Figure 3.

7.4 SUSTAINED CALIBRATION CAN REVERSE THE REGIME

Correct age does not immediately repair a pathological state, but it substantially improves the subsequent trajectory. We switch OWN to local-age reconstruction at update 4,800 and let projectors be recomputed from the resulting trajectory, with no healthy sequence imposed. Recovery is gradual: 800 updates later, Ψ is still 8.4 (17.0 without the switch); by update 10,000 it has fallen to 3.0, and perplexity is 42 instead of 330 (Appendix D.8). Calibrated reconstruction thus stops feeding the regime, and the gradient regime and training quality progressively improve endogenously. The mechanism is traced on layer-0 Q/K coordinates. Its consequences appear in R across all 56 GaLore matrices, but how far the mechanism itself propagates beyond these coordinates remains open.

8 DISCUSSION AND DESIGN IMPLICATIONS

The reset experiments change the question from what history must be preserved to what information must remain available for the next update. Joint-reset methods such as SPAM (Huang et al., 2025) rebuild both moments on a common support and keep the bound of Proposition 2 finite; retaining m while resetting v loses that bound, yet calibration is sufficient to recover carry-level quality in the tested runs. A denominator estimated from several current minibatches can also suppress the reconstruction shock almost as effectively as one built from historical gradients. The two states contain different histories but can provide similar normalization. What matters, therefore, is not necessarily reproducing the old buffer, but reconstructing enough of the function that buffer served. How much information is enough therefore depends on the state training has reached.

This distinction is important for interpreting ablations. Failure after removing a buffer late in training does not show that the buffer had to be preserved from the beginning; it shows that the buffer has become useful in the particular state the preceding trajectory created. Historical necessity and conditional usefulness are different questions. More generally, optimizer-state ablations can confound information removal with the calibration of the state that replaces it. If a rebuilt statistic is normalized inconsistently with what it has actually accumulated, the intervention changes both memory and effective update scale. This distinction applies beyond subspace refreshes whenever optimizer state is reset, reconstructed, or regrown.

Attention-logit growth has previously been associated with Transformer training instability (Wortsman et al., 2024). Here, it arises inside an optimizer-state reconstruction loop in which oversized updates reshape the gradient regime and make later reconstruction hazardous. Temporary amplitude control leaves a lasting imprint after its removal, and the future projector sequence governs persistence: past optimization is therefore remembered not only in m and v, but also in the parameters and in the adaptive choices made from the gradients those parameters produce. Reconstruction can change the state on which later reconstruction must operate. This is why the same local rule can be benign along one trajectory and hazardous along another.

The mechanistic route above is established through diagnostics localized to the instrumented layer-0 Q/K coordinates and selected heads. Its consequences are not confined to those coordinates: sustained attention normalization also reduces same-state reset sensitivity across all 56 GaLore matrices. The susceptibility grid does not establish that every severe run reaches failure through the same route. Since the fresh information needed for safe reconstruction depends on gradient intermittence, a natural direction is to let reconstruction windows respond to it, using short histories when the scale is easy to estimate and more observations when it is not. Our current implementation does not demonstrate optimizer-memory savings—resetting v at refresh does not remove the need to store v between refreshes—so this remains a design implication rather than a result.

9 CONCLUSION

A correct statistical age is sufficient to recover carry-level quality, but calibration alone does not make reconstruction state-invariant. In the GaLore settings we test, the complete pre-refresh history of Adam's second moment can be discarded repeatedly while recovering carry-level quality through 1B, provided the rebuilt state is calibrated to what it has actually accumulated. The full history of v therefore need not survive, but what must be available to reconstruct it safely depends on the gradient regime the trajectory has reached. The catastrophic failure of a naive reset therefore reflects more than lost history; it also reflects how the replacement state is constructed and how the resulting trajectory evolves.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing, coding, and selected mathematical derivations. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Xi Chen, Kaituo Feng, Changsheng Li, Xunhao Lai, Xiangyu Yue, Ye Yuan, and Guoren Wang. Fira: Can we achieve full-rank training of LLMs under low-rank constraint? In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.01623.

Arijit Das. Natural GaLore: Accelerating GaLore for memory-efficient LLM training and fine-tuning, 2024. URL https://arxiv.org/abs/2410.16029.

Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors, 2024. URL https://arxiv.org/abs/2402.03293.

Matan Haroush and Daniel Soudry. PLUMAGE: Probabilistic low-rank unbiased min-variance gradient estimator for efficient large model training, 2025. arXiv preprint arXiv:2505.18313.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2410.11289.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.06842.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DLJznSp6X3.

Sahar Rajabi, Nayeema Nonta, and Sirisha Rambhatla. Geometrically principled randomized optimization for efficient LLM training, 2025a. URL https://arxiv.org/abs/2510.01878.

Sahar Rajabi, Nayeema Nonta, and Sirisha Rambhatla. SubTrack++: Gradient subspace tracking for scalable LLM training. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2502.01586.

Thomas Robert, Mher Safaryan, Ionut-Vlad Modoranu, and Dan Alistarh. LDAdam: Adaptive optimization from low-dimensional gradient statistics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Zkp1GuHerF.

DiJia Su, Andrew Gu, Jane Xu, Yuandong Tian, and Jiawei Zhao. GaLore 2: Large-scale LLM pre-training by gradient low-rank projection. *arXiv preprint arXiv:2504.20437*, 2025. URL https://arxiv.org/abs/2504.20437.

Noel Thomas. No subspace to track: Non-identifiability and optimizer state in low-rank training, 2026. URL https://arxiv.org/abs/2607.05872.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie E. Everett, Alexander A. Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations*, 2024.

Jinqi Xiao, Shen Sang, Tiancheng Zhi, Jing Liu, Qing Yan, Linjie Luo, and Bo Yuan. COAP: Memory-efficient training with correlation-aware gradient projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.00071.

Qiao Xiao, Boqian Wu, Patrik Okanovic, Tomasz Sternal, Maurice van Keulen, Elena Mocanu, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Torsten Hoefler. Memory-efficient LLM training with dynamic sparsity: From stability to practical scaling. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2606.00888.

Zhenyu Zhang, Ajay Kumar Jaiswal, Lu Yin, Shiwei Liu, Jiawei Zhao, Yuandong Tian, and Zhangyang Wang. Q-GaLore: Quantized GaLore with INT4 projection and layer-adaptive low-rank gradients. In *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025. URL https://openreview.net/forum?id=KTAPk6c2hl.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A EXPERIMENTAL DETAILS AND INTERVENTIONS

Code availability An initial code release is available at `https://github.com/yt6yhywb4f-web/iclr2027-code`. The complete codebase will be released upon acceptance.

Configurations We train LLaMA-style decoders on C4/en with the `t5-base` tokenizer, following Zhao et al. (2024). Table 5 lists the GaLore configurations; no hyperparameter is retuned across interventions. Training uses 10% linear warmup followed by cosine decay to 10% of the peak learning rate. Low-rank projection is applied to the query, key, value, output, gate, up, and down projection matrices; all remaining parameters use full-rank AdamW. Validation perplexity is evaluated every 1,000 updates on at least 10M C4/en validation tokens.

Table 5: GaLore training configurations. Batch size is the global number of sequences per optimizer update.

	60M	130M	1B
Hidden size / layers / heads	512 / 8 / 8	768 / 12 / 12	2048 / 24 / 32
Sequence length / global batch	256 / 512	256 / 512	256 / 512
Optimizer updates	10K	20K	100K
Warmup updates	1K	2K	10K
GaLore rank r	128	256	1024
Peak learning rate	0.01	0.01	0.01
$(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-6}$, weight decay $= 0$, gradient clipping $= 0$			
BF16; GaLore scale $= 0.25$; `std` projection; refresh interval $T = 200$ unless stated otherwise			

Cross-method experiments use each optimizer's native configuration. At 60M, PLUMAGE and Fira use rank 128, scale 0.25, $T = 200$, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-6}$; Fira additionally uses residual-limiter coefficient $\gamma = 1.01$. LDAdam uses rank 128, $\rho = 0.908$, $(\beta_1, \beta_2) = (0.908, 0.99)$, $\epsilon = 10^{-8}$, power-iteration projection, and refreshes at every update. Realization identifiers are experiment-specific and are reported with the corresponding results in Appendix B; the C4 shuffle seed is fixed to 42.

Matched runs Within each matched realization, variants share initialization, data order, learning-rate schedule, and initial-projector construction, and differ only through the stated intervention. One-step counterfactuals branch from the exact same saved training state.

Native transitions and interventions At each refresh, the optimizer first performs its native state transition: GaLore retains projected moments across the new SVD basis; PLUMAGE realigns them; LDAdam applies its projection-aware transport and error feedback; and Fira refreshes its SVD projector while retaining its native residual path. Interventions are applied to the resulting state before the current gradient is accumulated:

	m^+	v^+	age used for v
`reset_m`	0	kept	global
`reset_v`	kept	0	global
`reset_all`	0	0	global
`local_v_age`	kept	0	local

The global optimizer counter is never reset. The first moment therefore always uses global bias correction, including under `reset_m` and `reset_all`; only v receives a local counter under `local_v_age`, with the first post-refresh update using local age one. Exact bias-correction and ϵ placement are given in Appendix C.

LDAdam LDAdam refreshes at every optimizer update. The asymmetry experiment therefore applies interventions at every update, so a reset second moment never accumulates more than the current projected gradient. The local-age calibration test instead applies a single reset at update 2,000, allowing v to rebuild afterward. LDAdam's error-feedback path remains active in both protocols.

Hardware Experiments were run on NVIDIA GB200, L40S, and H200 GPUs.

B COMPLETE QUANTITATIVE RESULTS

Ratios are relative to the matched no-reset baseline of the same realization.

Asymmetry and repeated local-age reconstruction Table 6 reports the completed endpoints that can be matched by method, scale, and realization. For PLUMAGE realization 123, the matched carry run ends at 42.40, markedly above the other two realizations. No execution failure was identified, so we report the paired absolute values rather than assigning a post-hoc cause to its lower local/carry ratio.

Table 6: Final validation perplexity for completed matched realizations. Repeated interventions are applied at every refresh. Fira runs use NVIDIA H200 GPUs. LDAdam is reported separately in Table 7.

Config	Realization	Baseline	`reset_m`	`reset_all`	`reset_v`	`local_v_age`
GaLore 60M	0	34.602	34.551	46.999	408.859	34.236
	2	34.589	34.560	47.102	404.434	34.087
	42	34.573	34.545	46.850	413.269	34.151
GaLore 130M	0	24.887	24.850	35.200	435.120	24.609
	1	24.942	24.910	35.450	438.905	24.603
	42	25.263	25.228	34.670	440.778	24.917
PLUMAGE 60M	0	34.56	34.52	48.30	232.66	33.692
	7	34.19	34.15	47.80	329.42	33.546
	123	42.40	42.35	58.20	498.44	33.523
Fira 60M	7	30.720	30.993	146.396	788.464	30.775
	123	30.772	31.080	134.838	588.813	30.837

LDAdam LDAdam refreshes at every update and is reported separately from Table 6. Table 7 gives the runs behind the LDAdam column of Table 2, with interventions applied at every refresh.

Table 7: LDAdam-60M final validation perplexity with interventions at every refresh (every update).

Realization	Baseline	`reset_m`	`reset_all`	`reset_v`
42	29.136	29.204	128.685	1391.431
123	29.102	29.088	126.435	878.583
314	29.186	29.312	130.267	788.532

LDAdam single-reset protocol The local-age calibration experiment instead uses a single reset at update 2,000. The first post-reset global/local update-norm ratio is $3.117940\times$, against an ϵ-aware prediction of $3.117698\times$; the ϵ-free prediction would be $10.0\times$. Under global age, training loss overshoots its pre-reset value by 0.177 and returns after 39 updates, whereas no overshoot is observed under local age.

Table 8: LDAdam-60M final validation perplexity after a single second-moment reset at update 2,000. Ratios to the matched baseline are in parentheses.

Realization	Baseline	Global-age reset	Local-age reset
42	30.410	30.608 (1.007×)	30.360 (0.998×)
123	29.048	29.134 (1.003×)	29.048 (1.000×)

Partial retention We additionally run complete GaLore-60M trajectories in which, at every refresh, the retained second moment is scaled as $v \leftarrow \gamma v$, with $\gamma \in \{0, 0.25, 0.5, 1\}$. All four trajectories use realization 42, $T = 200$, $\beta_2 = 0.999$, and a 10k-update horizon; $\gamma = 1$ is the matched carry reference. Positive retention remains close to carry: $\gamma = 0.5$ and 0.25 end at $1.0354\times$ and $1.0309\times$ the matched baseline, respectively, whereas complete erasure reaches $13.5494\times$.

Table 9: GaLore-60M complete repeated-retention trajectories. At every refresh, v is replaced by γv while the global optimizer age is retained. All runs use realization 42. These runs were produced on different hardware from those of Table 6, so the carry endpoint differs from the corresponding baseline there; comparisons are made within this table only.

γ	Carry PPL	Final PPL	Final / carry
1.00	35.7368	35.7368	1.0000
0.50	35.7368	37.0004	1.0354
0.25	35.7368	36.8426	1.0309
0.00	35.7368	484.2129	13.5494

C Derivations and proofs

We use the notation of Section 5. For one projected coordinate, the pre-refresh scale is σ^2, the post-refresh gradients satisfy $\mathbb{E}[g_t^2] = \sigma'^2$, and $\rho = \sigma^2/\sigma'^2$. The optimizer age immediately before refresh is s, and $t \geq 1$ counts updates since refresh.

C.1 Optimizer-state reconstruction

Second-moment reconstruction (Proposition 1). With $v_0^+ = \gamma v_{\mathrm{pre}}$, $\gamma \in [0, 1]$, unrolling the EMA gives

$$v_t = \beta_2^t v_0^+ + (1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} g_i^2.$$

Using $(1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} = 1 - \beta_2^t$, and, when retained state is present, $\mathbb{E}[v_{\mathrm{pre}}] = (1 - \beta_2^s)\sigma^2$, global bias correction yields

$$\frac{\mathbb{E}[\hat{v}_t]}{\sigma'^2} = \frac{\gamma\rho(1 - \beta_2^s)\beta_2^t + (1 - \beta_2^t)}{1 - \beta_2^{s+t}}, \tag{8}$$

which is Proposition 1. Only the common second-moment scale of the post-refresh gradients enters; no independence assumption is needed.

Complete reset For $\gamma = 0$ the retained term vanishes and equation 8 reduces to the factor $A_{\mathrm{eff}}(t; s)$ of equation 1, independently of ρ and of the calibration of v_{pre}. In the mature limit, $A_{\mathrm{eff}}(t) = (1 - \beta_2^t)^{-1/2}$ and $A_{\mathrm{peak}} = (1 - \beta_2)^{-1/2} = 31.62$ at $\beta_2 = 0.999$. For $a > 1$, $A_{\mathrm{eff}}(t) > a$ if and only if $t < t_a = \log(1 - a^{-2})/\log \beta_2$; at $\beta_2 = 0.999$, $t_5 \simeq 40.8$ and $t_2 \simeq 287.5$.

Local statistical age After a complete reset, $\hat{v}_t^{\mathrm{local}} = v_t/(1 - \beta_2^t)$, so $\mathbb{E}[\hat{v}_t^{\mathrm{local}}] = \sigma'^2$: the age-induced mean-field factor is exactly one. This is a statement about calibration, not about identity with the carried historical second moment.

Matched-state ratio When both branches share the raw post-gradient state and differ only through $c_G = 1 - \beta_2^{s+t}$ and $c_L = 1 - \beta_2^t$, their denominator ratio is

$$A_{\mathrm{age}}(t; s) = \sqrt{c_G/c_L}, \tag{9}$$

an algebraic identity requiring no stationarity assumption (for ϵ, see Appendix C.2). At $\beta_2 = 0.999$ and $t = 1$, the predicted ratios for $s = 200, 400, 600, 800$ are 13.497030, 18.179227, 21.257981 and 23.479779; the measured ratios are 13.494079, 18.176992, 21.257175 and 23.476155, all within 0.1%. At 1B, in a matched state at update 12,000 with identical projected gradients, first moments

and raw second moments between branches for all 168 GaLore projectors, the predicted ratio 31.6226801311 is reproduced with maximum relative error 2.3×10^{-16}. Over four matched refreshes, the median first post-refresh update norm is $25.322137\times$ the matched baseline norm under global age and $1.234180\times$ under local age; the median same-minibatch loss increases are $+3.6796875$ and 0.0, respectively.

Partial retention In the mature limit, equation 8 becomes

$$q_\gamma(t) = \gamma\rho\beta_2^t + (1-\beta_2^t) \;\geq\; \min(\gamma\rho, 1), \tag{10}$$

a convex combination, so any $\gamma\rho > 0$ provides a strictly positive floor. Relative to the carry branch, at fixed nonzero mean-field numerator,

$$\frac{U_\gamma}{U_{\text{carry}}} = \sqrt{\frac{\rho\beta_2^t + (1-\beta_2^t)}{\gamma\rho\beta_2^t + (1-\beta_2^t)}}. \tag{11}$$

At $t = 1$ and $\rho = 1$ this gives 1.41 and 2.00 for $\gamma = 0.5$ and 0.25, measured at 1.40 and 1.96.

Joint reconstruction (Eq. 2) Assume a locally constant post-refresh mean $\mathbb{E}[g_t] = \mu'$. After resetting m, $\mathbb{E}[m_t] = (1-\beta_1^t)\mu'$, and with the preserved global age $\mathbb{E}[\hat{m}_t] = \frac{1-\beta_1^t}{1-\beta_1^{s+t}}\mu'$. With a complete second-moment reset and the calibrated mean-field reference $U_{\text{ref}} = |\mu'|/\sigma'$, the reset-all amplification is

$$A_{\text{all}}(t;s) = \frac{1-\beta_1^t}{1-\beta_1^{s+t}}\, A_{\text{eff}}(t;s), \qquad \frac{A_{\text{all}}}{A_{\text{eff}}} < 1 \;\; (s > 0). \tag{12}$$

Interpreting A_{eff} as the corresponding reset-v gain additionally assumes $\mathbb{E}[m_{\text{pre}}] = (1-\beta_1^s)\mu'$. In the mature limit, $A_{\text{all}}(1) = (1-\beta_1)/\sqrt{1-\beta_2} = 3.16$ against $A_{\text{eff}}(1) = 31.62$. The measured norms at the matched median refresh are $12.1\times$ and $44.8\times$ the baseline, so the joint-reset update carries 27% of the reset-v norm rather than the mean-field 10%: the relation is exact between mean-field factors, not between realized norms, since the carried numerator of reset v need not equal μ'.

C.2 Effect of ϵ and bias-correction placement

Consider two branches with identical raw m and v, differing only in c_G and c_L, and write $A = A_{\text{age}} \geq 1$ (Eq. 9).

Correction inside the denominator If $u_B = \hat{m}/(\sqrt{v/c_B} + \epsilon)$, let $\nu = \sqrt{v/c_G}$, so that $\sqrt{v/c_L} = A\nu$. The coordinatewise magnitude ratio is

$$r_\epsilon = \frac{|u_G|}{|u_L|} = \frac{A\nu + \epsilon}{\nu + \epsilon} = 1 + (A-1)\frac{\nu}{\nu+\epsilon} \in [1, A], \tag{13}$$

tending to A when $\nu \gg \epsilon$ and to 1 when $\nu \ll \epsilon$: ϵ attenuates the observable ratio in coordinates of small scale.

Correction in the scalar step multiplier If instead $u_B = \sqrt{c_B}\, m/(\sqrt{v} + \epsilon)$, up to factors common to both branches, then $|u_G|/|u_L| = A$ exactly, since ϵ enters an identical raw denominator.

From coordinates to norms If $u_{G,j} = r_j u_{L,j}$, the squared norm ratio is a weighted average of the r_j^2, so $\min_j r_j \leq \|u_G\|/\|u_L\| \leq \max_j r_j$. With orthonormal projectors this carries over to the weight-space adaptive update; weight decay, residual paths and clipping are excluded from the reported norms.

Executed rules GaLore, standard PLUMAGE, and Fira apply second-moment bias correction through the scalar step multiplier,

$$\Delta\theta_{\text{adapt}} = -\eta\, \frac{\sqrt{1-\beta_2^{t_v}}}{1-\beta_1^{t_m}}\, \text{backproj}\left(\frac{m_t}{\sqrt{v_t} + \epsilon}\right).$$

Fira additionally applies its native scaled residual and limiter. LDAdam instead applies bias correction inside the denominator,

$$\Delta\theta_{\text{adapt}} = -\eta \,\text{backproj}\left[\frac{m_t/(1-\beta_1^{t_m})}{\sqrt{v_t/(1-\beta_2^{t_v})}+\epsilon}\right].$$

All predictions in the paper use the executed rule.

C.3 Repeated partial retention

Proposition 1 treats one intervention from calibrated state. Repeated retention acts on state already depleted by earlier interventions. Assume a fixed representation ($\rho = 1$), a constant scale $\mathbb{E}[g_t^2] = \sigma^2$, interval T, retention $\gamma \in [0, 1]$, and a mature global counter. Let $b = \beta_2^T$ and $q_k = \mathbb{E}[v_k^-]/\sigma^2$ be the normalized level just before the k-th intervention. Since $\mathbb{E}[v_k^+] = \gamma\,\mathbb{E}[v_k^-]$ and $\mathbb{E}[v_{k+1}^-] = b\,\mathbb{E}[v_k^+] + (1-b)\sigma^2$,

$$q_{k+1} = \gamma b\, q_k + (1-b), \qquad q_* = \frac{1-b}{1-\gamma b}, \qquad q_k - q_* = (\gamma b)^k (q_0 - q_*), \tag{14}$$

so the periodic regime is reached geometrically with ratio $\gamma\beta_2^T$. Within a stationary cycle, after t updates,

$$q_*(t) = \gamma\beta_2^t q_* + (1-\beta_2^t) = 1 - \frac{(1-\gamma)\beta_2^t}{1-\gamma\beta_2^T}, \qquad A_*(t) = q_*(t)^{-1/2}. \tag{15}$$

The limits recover $q_*(t) = 1-\beta_2^t$ for $\gamma = 0$ and $q_*(t) = 1$ for $\gamma = 1$.

Example At $\beta_2 = 0.999$, $T = 200$, $\gamma = 0.25$: $b \simeq 0.81865$, $q_* \simeq 0.228$, $q_*(1) \simeq 0.0580$ and $A_*(1) \simeq 4.15$, whereas the same retention applied once to calibrated state gives $q_{\text{iso}}(1) = 0.25\beta_2 + (1-\beta_2) \simeq 0.2508$ and $A_{\text{iso}}(1) \simeq 2.00$. Repeated depletion thus doubles the initial amplification, yet complete runs at $\gamma = 0.25$ remain within $1.04\times$ the baseline (Table 9). This fixed-representation calculation is not a quantitative prediction for GaLore, whose projector and gradient statistics change across refreshes.

C.4 Proof of Proposition 2

Write $m = \sum_i a_i g_i$ and $v = \sum_i b_i g_i^2$ with $\text{supp}(a) \subseteq \text{supp}(b)$. Weighted Cauchy–Schwarz gives

$$\Big(\sum_i a_i g_i\Big)^2 = \Big(\sum_i \frac{a_i}{\sqrt{b_i}}\sqrt{b_i}\, g_i\Big)^2 \le \Big(\sum_i \frac{a_i^2}{b_i}\Big)\Big(\sum_i b_i g_i^2\Big),$$

with equality for $g_i \propto a_i/b_i$ on $\text{supp}(b)$. Hence

$$\sup_{\{g_i\}: v>0} \frac{m^2}{v} = \sum_{i: b_i>0} \frac{a_i^2}{b_i}. \tag{16}$$

If some $a_j \neq 0$ with $b_j = 0$, that contribution moves the numerator without entering the denominator, and the ratio is unbounded. Let $A = \sum_i a_i$, $B = \sum_i b_i$, $p_i = a_i/A$ and $q_i = b_i/B$. Then $\sum_{i:b_i>0} a_i^2/b_i = (A^2/B)\sum_{i:q_i>0} p_i^2/q_i$, and, when $\text{supp}(p) \subseteq \text{supp}(q)$,

$$\sum_{i:q_i>0} \frac{p_i^2}{q_i} = 1 + \sum_{i:q_i>0} \frac{(p_i-q_i)^2}{q_i} = 1 + \chi^2(p\|q).$$

Including bias correction, $\hat{m} = m/c_1$ and $\hat{v} = v/c_2$, multiplies the supremum by c_2/c_1^2, so the total prefactor is $A^2 c_2/(Bc_1^2)$. This is Proposition 2.

Propagation under retention Let $C_t = \sum_{i:b_{t,i}>0} a_{t,i}^2/b_{t,i}$, with $C_t = +\infty$ when the support inclusion fails. If the raw moments are scaled as $m \leftarrow \alpha_t m$, $v \leftarrow \gamma_t v$ before the EMA update, every historical pair maps to $(\beta_1\alpha_t a_{t-1,i}, \beta_2\gamma_t b_{t-1,i})$ and the new gradient contributes $(1-\beta_1, 1-\beta_2)$. For $\gamma_t > 0$,

$$C_t = \frac{\beta_1^2}{\beta_2}\frac{\alpha_t^2}{\gamma_t}C_{t-1} + \frac{(1-\beta_1)^2}{1-\beta_2}, \tag{17}$$

so any positive retention keeps a finite C finite. If $\gamma_t = 0$ and $\alpha_t > 0$ (reset v), surviving numerator history has no denominator counterpart and $C_t = +\infty$; if $\alpha_t = \gamma_t = 0$ (reset all), $C_t = (1-\beta_1)^2/(1-\beta_2)$ after the new gradient.

C.5 A CONSTRUCTIVE MODEL OF REPRESENTATION FEEDBACK

This subsection gives an existence result, not a model of the language-model dynamics: an explicit loss, optimizer, and gradient-based representation selector in which the gain introduced by moment reconstruction changes the stability of future representation choices. It captures the feedback between reconstruction gain and representation selection, not the attention mechanism of Figure 3.

Loss and update rule Let $w = (x, y)^\top \in \mathbb{R}^2$ and

$$\ell(x, y) = \tfrac{\kappa}{2}x^2 + y, \qquad \nabla\ell = (\kappa x, 1)^\top, \qquad \kappa > 0, \tag{18}$$

with decoupled weight decay $\lambda > 0$, step size $\eta > 0$, $\delta = 1 - \eta\lambda \in (0, 1)$ and $d = \delta^T$ for refresh period $T \geq 1$. At the start of cycle k, with $w_k = (x_k, y_k)^\top$, select and hold fixed for T updates

$$\theta_k = \alpha_{\text{sel}}\theta_{k-1} + (1-\alpha_{\text{sel}})\arctan(\kappa x_k), \qquad u_k = (\sin\theta_k, \cos\theta_k)^\top, \qquad 0 \leq \alpha_{\text{sel}} < 1. \tag{19}$$

For $\alpha_{\text{sel}} = 0$, u_k is the normalized loss gradient, i.e. the leading left singular vector of the 2×1 gradient matrix with positive second component; $\alpha_{\text{sel}} > 0$ is a specific angular smoothing, not the selector of any particular optimizer. Let μ_k be the scalar first moment before the refresh. Every branch sets $v_{k,0} = 0$; branches G and L carry $m_{k,0} = \mu_k$, and branch J sets $m_{k,0} = 0$. From $w_{k,0} = w_k$, for $t = 1, \ldots, T$,

$$g_{k,t} = u_k^\top \nabla\ell(w_{k,t-1}), \quad m_{k,t} = \beta_1 m_{k,t-1} + (1-\beta_1)g_{k,t}, \quad v_{k,t} = \beta_2 v_{k,t-1} + (1-\beta_2)g_{k,t}^2, \tag{20}$$

$$a_{k,t}^B = \frac{m_{k,t}}{\sqrt{v_{k,t}/c_t^B}}, \qquad w_{k,t} = \delta w_{k,t-1} - \eta a_{k,t}^B u_k, \qquad c_t^G = c_t^J = 1, \quad c_t^L = 1 - \beta_2^t, \tag{21}$$

with $w_{k+1} = w_{k,T}$ and $\mu_{k+1} = m_{k,T}$. G is reset v with global age, L reset v with local age, and J reset of both moments with global age. Global correction factors are set to one (mature-age idealization) and $\epsilon = 0$; the local factor is exact, and J applies no local correction to its rebuilt first moment.

Reference orbit At $x_k = \theta_{k-1} = 0$, $u_k = e_2$ and every projected gradient equals one. Then $v_{k,t} = 1 - \beta_2^t$ in all branches, $m_{k,t} = 1$ for G, L and $m_{k,t} = 1 - \beta_1^t$ for J, giving the exact scalar update profiles

$$q_t^G = \frac{1}{\sqrt{1-\beta_2^t}}, \qquad q_t^L = 1, \qquad q_t^J = \frac{1-\beta_1^t}{\sqrt{1-\beta_2^t}}, \qquad H_B = \eta\sum_{t=1}^{T}\delta^{T-t}q_t^B. \tag{22}$$

The longitudinal coordinate obeys $y_{k+1} = dy_k - H_B$, with unique cycle-boundary fixed point $y_B^* = -H_B/(1-d)$. With $x = \theta = 0$ and the profiles above, this defines a T-periodic reference orbit whose representation and transverse coordinate are shared by all branches.

Linearization Take $z_k = (x_k, \theta_{k-1}, y_k, \mu_k)^\top$, with reference $z_B^* = (0, 0, y_B^*, \mu_B^*)^\top$, $\mu_G^* = \mu_L^* = 1$, $\mu_J^* = 1 - \beta_1^T$, and $\delta z_k = z_k - z_B^*$. The second moment is discarded at every refresh and need not enter the state; all within-cycle second moments are positive near the orbit, so the cycle map is smooth there. The selector expands as $\theta_k = \alpha_{\text{sel}}\theta_{k-1} + \chi x_k + O(\|\delta z_k\|^3)$ with $\chi = (1-\alpha_{\text{sel}})\kappa > 0$, and the projected gradient as $g_{k,t} = \kappa x_{k,t-1}\sin\theta_k + \cos\theta_k = 1 + O(\|\delta z_k\|^2)$. Transverse perturbations thus affect the reconstructed moments only at second order; a first-moment perturbation changes $a_{k,t}^B$ at first order, but enters the transverse update multiplied by $\sin\theta_k$. Hence

$x_{k,t} = \delta x_{k,t-1} - \eta q_t^B \theta_k + O(\|\delta z_k\|^2)$ and, over a cycle, $x_{k+1} = dx_k - H_B\theta_k + O(\|\delta z_k\|^2)$. The transverse block and full cycle Jacobian are

$$J_B = \begin{pmatrix} d - H_B\chi & -\alpha_{\mathrm{sel}} H_B \\ \chi & \alpha_{\mathrm{sel}} \end{pmatrix}, \qquad \mathcal{J}_B = \begin{pmatrix} d - H_B\chi & -\alpha_{\mathrm{sel}} H_B & 0 & 0 \\ \chi & \alpha_{\mathrm{sel}} & 0 & 0 \\ 0 & 0 & d & -C_B \\ 0 & 0 & 0 & r_B \end{pmatrix}, \tag{23}$$

with $r_G = r_L = \beta_1^T$, $C_B = \eta \sum_{t=1}^{T} \delta^{T-t} \beta_1^t q_t^B$ for $B = G, L$, and $r_J = C_J = 0$. The two extra eigenvalues, d and r_B, lie strictly inside the unit disk, so stability is governed by J_B.

Proposition 3 (Representation feedback)**.** *In the model above, the reference orbit of branch B is locally exponentially stable if $H_B\chi < (1+d)(1+\alpha_{\mathrm{sel}})$ and unstable if $H_B\chi > (1+d)(1+\alpha_{\mathrm{sel}})$. Under replay of the reference representation sequence, it is stable in every branch. There exist curvatures κ for which, with the same selector, step size and refresh period, G is unstable while L and J are stable.*

Proof. Let $\alpha = \alpha_{\mathrm{sel}}$ and $h = H_B\chi$. Then $\operatorname{tr} J_B = d + \alpha - h$ and $\det J_B = d\alpha$. Both roots of $r^2 - \tau r + D$ lie strictly inside the unit disk if and only if $1 - \tau + D > 0$, $1 + \tau + D > 0$ and $1 - D > 0$, here $(1-d)(1-\alpha) + h > 0$, $(1+d)(1+\alpha) - h > 0$ and $1 - d\alpha > 0$. Under $0 < d < 1$, $0 \le \alpha < 1$ and $h > 0$, the first and third hold automatically, which gives the stability condition. If it is reversed, the characteristic polynomial is negative at $r = -1$ and tends to $+\infty$ as $r \to -\infty$, so it has a real root below -1. (At equality, the eigenvalues are -1 and $-d\alpha$, and linearization is inconclusive.) Under replay, $u_k = e_2$ independently of the trajectory, so $x_{k+1} = dx_k$ exactly and the remaining multipliers are d and r_B. Finally, since $0 < \beta_1, \beta_2 < 1$, $0 < q_t^L < q_t^G$ and $0 < q_t^J < q_t^G$ for every t; the weights in H_B are positive, so $H_L, H_J < H_G$. Any $\chi \in \big((1+d)(1+\alpha)/H_G,\ (1+d)(1+\alpha)/\max(H_L, H_J)\big)$, a nonempty interval, separates the branches and is realized by the common curvature $\kappa = \chi/(1-\alpha)$. □

Instance With $T = 200$, $\eta = 10^{-3}$, $\lambda = 1$, $\kappa = 2.6$, $\alpha_{\mathrm{sel}} = 0$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (parameters of the construction, not measurements), $d \simeq 0.8186$, $H_G \simeq 0.7612$, $H_L \simeq 0.1814$ and $H_J \simeq 0.6562$. The nonzero transverse eigenvalue $d - H_B\chi$ is -1.160 for G, 0.347 for L and -0.888 for J: only G is unstable, and all three are stable under replay.

Scope The construction shows that a reconstruction-dependent gain can change the stability of a gradient-driven representation-selection loop, and that replay can remove the destabilizing feedback. It does not identify the unstable mode of the language-model experiments or establish a finite-counter stability theorem for Adam.

C.6 LDAdam single-reset check

LDAdam uses $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$ in this experiment, and a single reset of v is applied at update 2,000. The ϵ-free global/local age factor, equation 9, is approximately 10.0. Evaluating the executed ϵ-aware rule (Appendix C.2) on the stored matched seed-42 state predicts an update-norm ratio of 3.117698, against 3.117940 measured in the instrumented run. Under global age, training loss overshoots its pre-reset value by 0.177 and returns after 39 updates; no corresponding overshoot is observed under local age. Final perplexities of the complete runs are reported in Table 8.

D Mechanism details

Unless stated otherwise, diagnostics use the reference GaLore–LLaMA-60M configuration at $(\beta_2, T) = (0.999, 200)$. OWN resets v with the global optimizer age and recomputes projectors from its own trajectory; REPLAY uses the same reset rule with projectors from a matched healthy trajectory.

D.1 Diagnostic quantities

Reset sensitivity R_k (Eq. 6) compares, from the same state, a local-age reset of v with the native no-reset update.

Table 10: Same-state reset sensitivity along the reference OWN trajectory. The update norms defining R_k are realized BF16 parameter displacements over the 56 GaLore matrices; Ψ_k is computed on projected layer-0 Q/K coordinates.

Update	No-reset norm	Local-age reset norm	R_k	Ψ_k
2,000	8.4	2.1	0.25	1.3
4,000	7.8	24.2	3.10	8.6
4,800	7.5	81.0	10.8	14.8

Historical scale and fixed-momentum mismatch The historical reference uses the 200 gradients immediately preceding the probe at checkpoint k, each produced at its original historical parameter state and re-expressed in the current projector P_k:

$$\tilde{g}_i = \Pi_{P_k}(G_i), \qquad v_k^H = (1-\beta_2)\sum_{i=k-200}^{k-1} \beta_2^{k-1-i}\tilde{g}_i^{\odot 2}, \qquad s_{k,j} = \sqrt{\frac{v_{k,j}^H}{1-\beta_2^{200}}}. \tag{24}$$

Thus v_k^H contains only gradients preceding the current probe gradient g_k.

Fixed-momentum diagnostics For both F_τ and Ψ, all coordinates below are restricted to projected layer-0 Q/K coordinates. Let m_k^- denote the native first moment before incorporation of the current gradient. We define

$$F_\tau(k) = 100\,\frac{\sum_j (m_{k,j}^-)^2 \mathbf{1}[|g_{k,j}| \le \tau s_{k,j}]}{\sum_j (m_{k,j}^-)^2}, \qquad \tau = 0.10, \tag{25}$$

with $\tau = 0.05, 0.20$ used as sensitivity checks.

For Ψ, both normalized updates use the fixed numerator $n_k = m_k^-/(1-\beta_1^k)$. With

$$d(v,c) = \frac{\sqrt{v}+\epsilon}{\sqrt{c}}, \qquad v_k^F = (1-\beta_2) g_k^{\odot 2},$$

we define

$$\Psi_k = \frac{\left\|\mathcal{B}_{P_k}\left(n_k \oslash d(v_k^F, 1-\beta_2)\right)\right\|_F}{\left\|\mathcal{B}_{P_k}\left(n_k \oslash d(v_k^H, 1-\beta_2^{200})\right)\right\|_F}. \tag{26}$$

$\mathcal{B}_{P_k}$ includes GaLore backprojection and scaling; weight decay is excluded. Ψ is a fixed-momentum normalization diagnostic, not the complete native Adam update.

Trajectory intermittency and attention For the natural-trajectory diagnostics, we use 64 prespecified observations from the preceding 200-update window. Each gradient is the gradient actually produced at its historical parameter state and is re-expressed in the checkpoint projector P_k. With

$$E_i = \sum_{j \in Q/K0} \tilde{g}_{i,j}^2,$$

Top-5 is the fraction of $\sum_i E_i$ carried by the five largest E_i. Median $|g|/s$ is computed over the same observations and Q/K0 coordinates. Logit scale is centered RMS, and normalized attention entropy is $-\sum_k p_k \log p_k / \log |\mathcal{K}_q|$, averaged over queries, minibatches, and selected heads. In the main text, Top-5 is called gradient concentration and $F_{0.10}$ low-gradient momentum.

D.2 Replication across realizations

All three realizations reproduce the joint increase in mismatch, gradient concentration, low-gradient momentum, and logit scale under late OWN (Table 11). The main interventions also replicate in direction and approximate magnitude (Table 12).

Table 11: Diagnostic transition across three matched GaLore-60M realizations. Low-gradient momentum denotes $F_{0.10}$ (Eq. 25); Top-5 is gradient concentration.

Seed	State	Ψ	Top-5	median $\lvert g\rvert/s$	low-grad. mom.	logit scale	entropy
123	OWN 2.0k	1.3	38%	.58	10%	2.4	.43
123	OWN 4.8k	14.8	90%	.035	78%	10.2	.07
123	REPLAY 4.8k	1.5	41%	.61	11%	3.0	.39
51	OWN 2.0k	1.4	35%	.54	12%	2.6	.41
51	OWN 4.8k	12.6	85%	.052	69%	8.9	.10
51	REPLAY 4.8k	1.7	44%	.56	13%	3.2	.36
34	OWN 2.0k	1.2	42%	.64	9%	2.3	.45
34	OWN 4.8k	16.1	88%	.030	76%	11.0	.06
34	REPLAY 4.8k	1.3	38%	.68	9%	2.8	.42

Table 12: Replication of the mechanistic interventions across three GaLore-60M realizations, at update 4,800 unless noted. *Intervention* is, per row group: frozen-state normalization of the selected OWN logits (Appendix D.3); ATTN-NORM (Appendix D.4); PROTECTED (Appendix D.6); LATE-LOCAL (Appendix D.8).

	Seed 123		Seed 51		Seed 34	
	OWN	Interv.	OWN	Interv.	OWN	Interv.
Frozen attn.: Ψ	14.8	5.4	12.6	5.1	16.1	6.2
Frozen attn.: low-grad. mom.	78%	43%	69%	40%	76%	46%
Sustained attn.: R	10.8	1.4	8.9	1.6	12.3	1.8
Sustained attn.: Ψ	14.8	2.8	12.6	3.1	16.1	3.4
Protection: $\Psi_{5.6k}$	17.0	2.0	14.7	2.4	18.5	2.7
Late local-age: Ψ_{10k}	19.4	3.0	16.8	3.5	21.2	3.8
Late local-age: PPL_{10k}	330	42	276	45	382	47

D.3 ATTENTION-LOGIT INTERVENTIONS

At update 4,000, layer-0 Q/K heads are ranked by

$$E_h = \|\Delta W_{Q,h}\|_F^2 + \|\Delta W_{K,h}\|_F^2.$$

The smallest set covering at least 80% of the energy is retained ($h3, h7, h1$; 85% total) and then fixed. For selected heads,

$$L'_{h,qk} = \mu_{h,q} + \alpha_h(L_{h,qk} - \mu_{h,q}), \tag{27}$$

with $(\alpha_{h3}, \alpha_{h7}, \alpha_{h1}) = (0.34, 0.38, 0.42)$, obtained on a separate calibration bank as the REPLAY-to-OWN ratio of centered-logit RMS for each selected head. This calibration bank is distinct from the frozen-state evaluation minibatches used below. The reciprocal intervention uses $1/\alpha_h$.

At update 4,800, native and intervened probes share parameters, m^-, projector, minibatches, and historical gradient reference. The α_h are detached constants, but the transformed logits remain in the computational graph and the full gradient is recomputed; the SVD is not recomputed. We use two distinct same-state evaluations. First, we recompute the current gradient under each intervention and evaluate Ψ and $F_{0.10}$ using the native first moment and fixed historical scale reference. Second, to measure gradient intermittency locally at the checkpoint rather than along the preceding trajectory, we rerun the same 64 held-out minibatches at the fixed update-4,800 parameters and projector under the native and intervened forwards. Frozen-state Top-5 and median $\lvert g\rvert/s$ are computed from these recomputed gradients. The historical gradient bank defining s and the denominator of Ψ remains unchanged, as does the native first moment m^-. These frozen-state quantities are distinct from the natural-trajectory Top-5 and median $\lvert g\rvert/s$ reported in Section D.2.

The interventions move the diagnostics in opposite directions, establishing a bidirectional local effect without implying full reproduction of the opposite state.

Table 13: Bidirectional frozen-state attention interventions at update 4,800. Top-5 and median $|g|/s$ are recomputed on the same 64 held-out minibatches; historical references are fixed across each comparison.

Diagnostic	OWN	OWN norm.	REPLAY	REPLAY OWN-like
Logit scale	10.2	4.0	3.0	8.6
Entropy	.07	.28	.39	.13
Frozen Top-5	86%	68%	38%	74%
Frozen median $\|g\|/s$	.052	.18	.57	.15
Low-grad. momentum	78%	43%	11%	52%
Ψ	14.8	5.4	1.5	6.3

D.4 SUSTAINED ATTENTION INTERVENTION

As a retrospective mechanistic intervention, ATTN-NORM ("Attn. norm." in Table 4) starts from a common checkpoint near update 2,200 and thereafter applies the logit transformation from Section D.3 to the selected heads, while using the same data, learning-rate schedule, initial state, and archived projector sequence as OWN.

Table 14: Sustained attention intervention at update 4,800.

Diagnostic	OWN	ATTN-NORM
Ψ	14.8	2.8
Top-5	90%	53%
Low-grad. momentum	78%	25%
Logit scale	10.2	4.2
Entropy	.07	.27
R (local-age / no-reset)	10.8	1.4
PPL	72	39

Healthy carry has PPL approximately 32 at the same stage.

D.5 SAME-STATE AMPLITUDE INTERVENTIONS

At the late OWN state, all branches share parameters, optimizer state, projector, and evaluated minibatch. The 32-minibatch denominator is estimated from auxiliary gradients at the same frozen parameters and projector; the recent-gradient denominator uses Eq. 24. Neither restores historical v.

Table 15: Same-state interventions at the late OWN state. Norms are relative to the ordinary local-age-reset update.

Intervention	Relative adaptive norm	Loss increase
Local-age reset	1.00	$+0.44$
32-minibatch denominator	0.07	$+0.02$
Recent-gradient denominator	0.06	$+0.02$
Original direction, norm matched to 0.06	0.06	$+0.03$
Zero retained m	0.03	$+0.01$

Matching only the recent-gradient update norm gives nearly the same loss change as using the recent-gradient denominator, indicating that the acute damage at this state is primarily an amplitude effect.

D.6 TEMPORARY AMPLITUDE PROTECTION

PROTECTED preserves each matrix's adaptive-update direction but matches its norm to that produced by the auxiliary recent-gradient normalization,

$$u^{\mathrm{prot}} = \frac{\|u^{\mathrm{shadow}}\|}{\|u^{\mathrm{OWN}}\|} u^{\mathrm{OWN}}. \tag{28}$$

Protection is applied from updates 2,000 through 4,000 and removed at 4,200.

Table 16: Effect of temporary protection; all protection has been removed by updates 4,400 and 5,600.

Branch	Ψ_{4000}	Ψ_{4400}	Ψ_{5600}	Logit scale$_{5600}$
OWN	8.6	11.9	17.0	11.1
Temporary recent-v	1.5	1.6	1.7	3.3
Temporary amplitude protection	1.7	1.8	2.0	3.7

The protection remains effective after removal, showing that preventing the earlier large displacements changes the subsequent state.

D.7 Current versus future projector interventions

All branches start from the same OWN state immediately before the refresh at update 4,800, sharing parameters, optimizer state, global counter, and next minibatch. OWN uses its native projector at 4,800 and subsequently recomputes projectors from its own trajectory. In the single-projector swap, only the projector at the 4,800 refresh is replaced by the matched healthy projector; later projectors are again recomputed endogenously from that branch. In the future-sequence splice, the matched healthy projectors are imposed from the 4,800 refresh onward (4,800, 5,000, 5,200, ...).

Table 17: Projector interventions branched immediately before the update-4,800 refresh and evaluated 800 updates later, at update 5,600.

Condition	Reset-boundary norm at 5,600
OWN continuation	228
Single healthy-projector swap at 4,800	176
Healthy future-sequence splice from 4,800	21

A single projector replacement provides limited relief, whereas controlling the subsequent projector sequence strongly changes the later reset-boundary norm. Because the interventions differ in duration, this comparison does not quantify the relative importance of the current and future projectors; it shows that the future representation sequence can causally alter persistence of the pathological regime.

D.8 Late switch to calibrated reconstruction

Starting from the exact OWN state at update 4,800, LATE-LOCAL continues to reset v at every refresh but switches to local v age. Both branches recompute all future projectors from their own trajectories; no healthy sequence is replayed.

Table 18: Late switch to local-age reconstruction from the pathological OWN state.

Step	Branch	Ψ	Top-5	Low-grad. mom.	Logit scale	Entropy	PPL
4.8k	common	14.8	90%	78%	10.2	.07	72
5.6k	OWN	17.0	92%	82%	11.1	.05	118
5.6k	LATE-LOCAL	8.4	81%	61%	8.0	.13	76
7.0k	OWN	18.3	91%	84%	11.7	.05	210
7.0k	LATE-LOCAL	4.9	67%	41%	5.9	.23	58
10k	OWN	19.4	92%	86%	12.0	.04	330
10k	LATE-LOCAL	3.0	55%	27%	4.4	.29	42

Recovery is not immediate, but all diagnostics improve progressively under LATE-LOCAL while its future projectors remain endogenous.