

Compressing Only What Is Compressible: Memory-Efficient Multiscale Muon via Selective Low-Rank Projection

Lianhai Ren^{1,2*} Yeyun Gong^{2†}

¹Department of Mathematics, National University of Singapore

²Microsoft Research Asia

lianhai@nus.edu yegong@microsoft.com

Abstract

Multiscale Muon maintains K momentum EMAs at different decay rates β_k , doubling or tripling Muon’s optimizer-state memory. We show that each EMA is a first-order low-pass filter and Multiscale Muon is therefore a *filter bank* on the gradient stream. Under this view its empirical gain over single- β Muon is a direct consequence of multi-band coverage, and an asymmetric compressibility follows from the bandwidth: the narrow-band (high- β) buffer is dominated by the slow, low-rank gradient component and is safe to project, while the wide-band (low- β) buffer carries high-frequency content whose effective rank is large and cannot be compressed without hurting convergence. We formalize this asymmetry and instantiate it in S-MUG (Selective Multiscale Muon with GaLore), applying GaLore only to high- β buffers. On 201 M GPT-2 pretraining, S-MUG ($K=2$) matches full-rank Multiscale Muon in perplexity (21.66 vs. 21.69) at $1.25\times$ Muon’s optimizer memory instead of $2.00\times$; the result transfers to 756 M, as well as 1 B and 3 B GQA scale-transfer pretraining, and composes cleanly with other single-buffer Muon improvements like NorMuon and TEON.

1 Introduction

Multiscale Muon (Pappone and Selvaraj, 2025) improves over single- β Muon (Jordan et al., 2024) by maintaining K independent EMA momentum buffers at distinct decay rates $\beta_1 < \dots < \beta_K$, capturing gradient history at multiple temporal scales simultaneously and consistently improving perplexity. The cost is direct: storing K full-rank momentum matrices multiplies optimizer-state memory by K . Naively applying GaLore (Zhao et al., 2024) to all K buffers fails: low-rank projection across

every buffer destroys convergence, leaving practitioners with a hard choice between quality and memory.

The key observation is that the K buffers are not equally compressible. Each EMA update acts as a first-order low-pass filter on the gradient stream (Section 3.1): a narrow-band (high- β) buffer retains only the slowly-varying, low-frequency gradient component, which concentrates in a low-rank subspace. A wide-band (small- β) buffer tracks the full gradient spectrum, including transient high-frequency directions that span a near-full-rank subspace. We formalise this bandwidth–rank monotonicity in Proposition 1 and confirm it empirically: applying GaLore to the wide-band buffer consistently hurts convergence, while applying it only to the narrow-band buffer preserves quality (Table 7).

This asymmetry yields a simple design rule: *compress only the narrow-band buffer*. We instantiate it in S-MUG (Selective Multiscale Muon with GaLore), which retains the wide-band β_1 buffer at full rank while applying GaLore only to the narrow-band β_2 buffer, recovering 75% of Multiscale Muon’s added optimizer-state overhead at no average quality cost. On 200 M GPT-2 with $n=3$ seeds, $K=2$ S-MUG matches full-rank Multiscale Muon in perplexity (21.66 ± 0.05 vs. 21.69 ± 0.03) using $1.25\times$ Muon’s optimizer memory (incl. the persistent GaLore basis) vs. $2.00\times$; results transfer to 756 M, as well as 1 B and 3 B GQA scale-transfer pretraining pipelines.

Our contributions are: (i) a filter-bank view of multiscale momentum (Section 3), casting (β_k, s_k) as a frequency-domain design choice rather than an unstructured set of hyperparameters; (ii) bandwidth–stable rank monotonicity (Proposition 1), supported by singular-value spectra (Appendix C) and the bandwidth-dependent shift of λ^* (Section 5.2); and (iii) the S-MUG algorithm (Sections 4–5), which reduces Multiscale Muon’s added optimizer-state overhead by $4\times$ at $K=2$, matches

* Work done during an internship at Microsoft Research Asia.

† Corresponding author: yegong@microsoft.com.

full-rank quality across 201 M–3 B scales, and composes cleanly with single-buffer Muon variants like NorMuon and TEON.

2 Background and Setup

2.1 Optimization Setup

We minimise an empirical loss $\mathcal{L}(W) = \mathbb{E}_\xi[\ell(W; \xi)]$ for a single weight matrix $W \in \mathbb{R}^{m \times n}$ (block-coordinate analysis applies independently to each tensor). At step t the optimizer observes a stochastic gradient $G^t = \nabla_W \ell(W^t; \xi^t) \in \mathbb{R}^{m \times n}$ and returns an update $W^{t+1} = W^t - \eta \Delta W^t$. We treat $\{G^t\}$ as a discrete-time matrix-valued signal; this makes the filter-bank view of momentum (Section 3) precise.

2.2 Muon

Muon (Jordan et al., 2024) maintains a single EMA momentum buffer $M^t = \beta M^{t-1} + (1 - \beta) G^t$ (the $K=1$ specialisation of Eq. (1); equivalent up to a constant rescaling to the classical heavy-ball form, since Newton-Schulz normalises by Frobenius norm) and applies Newton-Schulz (NS) orthogonalization to obtain $\hat{U}^t = \text{NS}(M^t) \approx UV^\top$ (where $M^t = U\Sigma V^\top$), with update $W^{t+1} = W^t - \eta \hat{U}^t$. Because $\hat{U}^t$ has approximately unit spectral norm, this realises steepest descent under the spectral-norm trust region $\|\Delta W\|_2 \leq \eta$ (Jordan et al., 2024; Bernstein and Newhouse, 2024). The quintic NS iteration $X_{i+1} = aX_i + bX_iX_i^\top X_i + c(X_iX_i^\top)^2 X_i$ with the standard Muon scalars $(a, b, c) = (3.4375, -4.375, 2.1875)$ (Jordan et al., 2024), started from $X_0 = M^t / \|M^t\|_F$, is a fixed-step orthogonalization heuristic that closely approximates the polar factor $UV^\top$ in q steps without an explicit SVD and extends to non-square buffers via the right-grouped form $X \leftarrow X(aI + bX^\top X + c(X^\top X)^2)$ (Appendix B).

2.3 Multiscale Muon

Multiscale Muon (Pappone and Selvaraj, 2025) maintains K independent EMA buffers with $0 \leq \beta_1 < \dots < \beta_K < 1$:

$$\begin{aligned} M_k^t &= \beta_k M_k^{t-1} + (1 - \beta_k) G^t, \\ W^{t+1} &= W^t - \eta \sum_{k=1}^K s_k \text{NS}(M_k^t), \\ s_k &\propto (1 - \beta_k), \quad \sum_k s_k = 1. \end{aligned} \quad (1)$$

The scale s_k is applied *after* NS, which normalises by Frobenius norm and erases any pre-NS scalar,

so s_k acts strictly as a post-NS band gain. The default choice $s_k \propto (1 - \beta_k)$ (with $\sum_k s_k = 1$ matching Muon’s spectral-norm trust region) allocates more weight to wide-band channels carrying transient short-horizon corrections; the alternative *free- λ* parametrisation $s_1 = 1 - \lambda$, $s_2 = \lambda$ used in our λ -sweep (Section 5.2) places weight λ on the narrow-band buffer and decouples the band-gain ratio from β . Total momentum memory is $K \cdot mn$, the central cost S-MUG addresses.

2.4 GaLore

GaLore (Zhao et al., 2024) projects the gradient onto a slowly-rotating low-rank subspace before forming momentum: with $V \in \mathbb{R}^{n \times r}$ (top- r right singular vectors of a recent gradient, refreshed every T steps with M_r reset), $M_r^t = \beta M_r^{t-1} + G^t V \in \mathbb{R}^{m \times r}$ and the full-space update is reconstructed as $M_r^t V^\top$. Per-buffer memory drops from mn to mr .

3 Multiscale Momentum as a Filter Bank

In this section we develop a frequency-domain analysis of multiscale momentum that provides theoretical grounding for the selective compression strategy of S-MUG.

3.1 Transfer Function and Bandwidth of EMA Momentum

Initialising $M^0 = 0$ and adopting the unscaled convention $M^t = \beta M^{t-1} + G^t$ (equivalent up to a fixed $(1 - \beta)$ rescaling that we restore in our implementation; cf. Eq. (1)), the update unrolls as

$$M^t = \sum_{s=0}^{t-1} \beta^s G^{t-s}, \quad (2)$$

showing that M^t is a weighted sum of the gradient history with exponentially decaying coefficients (we reserve τ for the time constant $\tau = 1/(1 - \beta)$ throughout). Moving to the frequency domain via the z -transform (Oppenheim and Schaffer, 1999), the corresponding transfer function is

$$H(z; \beta) = \frac{1}{1 - \beta z^{-1}}, \quad (3)$$

a first-order IIR low-pass filter (Oppenheim and Schaffer, 1999) with a single pole at $z = \beta$. Its frequency response has magnitude

$$|H(e^{j\omega}; \beta)| = \frac{1}{\sqrt{1 - 2\beta \cos \omega + \beta^2}}, \quad (4)$$

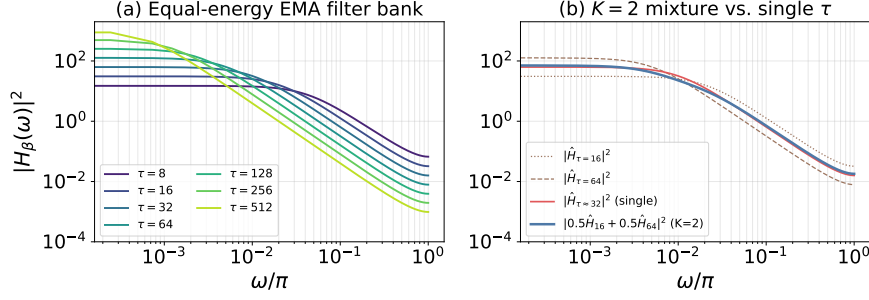


Figure 1: Multiscale Muon as a subband filter-bank synthesis system. **(a)** EMA filter bank with the equal-energy normalisation $\hat{H}_\beta(\omega) = \sqrt{1 - \beta^2} / (1 - \beta e^{-j\omega})$, $\int |\hat{H}_\beta|^2 d\omega / 2\pi = 1$, chosen so that the steady-state update has β -independent norm after per-parameter rescaling. **(b)** The $K=2$ mixture $|0.5\hat{H}_{\tau=16} + 0.5\hat{H}_{\tau=64}|^2$ has a flatter passband than the geometric-mean single filter $|\hat{H}_{\tau \approx 32}|^2$ at the same total energy.

with effective time constant $\tau \approx 1/(1 - \beta)$ steps. Choosing β is therefore choosing a *temporal bandwidth*: small β ($\tau \approx 1$, wide-band) tracks the instantaneous gradient, large β ($\tau \gg 1$, narrow-band) retains only the slowly-varying low-frequency component.

3.2 Multiscale Muon as Subband Filter-Bank Synthesis

Multiscale Muon thus applies K parallel low-pass filters $\{H(\cdot; \beta_k)\}$ to the gradient stream (a *filter bank*; Figure 1) and the combined update $\Delta W^t = \sum_k s_k \text{NS}(M_k^t)$ performs *subband synthesis* (Vaidyanathan, 1993): each filtered gradient is independently orthogonalised and recombined with a band gain s_k . A single EMA forces a sharp bandwidth/noise trade-off: a small- β filter passes too much high-frequency noise, while a large- β filter is too smooth to track the useful low-to-mid-frequency component of the gradient. This is the band whose empirical importance is fit by the score $g(\omega)$ in Section 5.2 (positive on low frequencies, sign-flipped on high frequencies). A filter bank with two or more β_k values covers this useful band more flatly than any single EMA can.

3.3 Bandwidth–Rank Monotonicity and Compressibility

GaLore (Zhao et al., 2024) observes that LLM gradient matrices approximately lie in a slowly rotating low-rank subspace, naturally a statement about the *low-frequency* component of the gradient sequence. We formalise the consequence for momentum compressibility:

Proposition 1 (Bandwidth–rank monotonicity; informal). *Let $M_k^t = \sum_{s \leq t} \beta_k^{t-s} G^s$ with $\beta_k \in [0, 1)$.*

Suppose the gradient stream admits a decomposition $G^s = G^{s,\parallel} + G^{s,\perp}$, where $G^{s,\parallel}$ has approximately stationary, slowly-rotating row-space (the GaLore subspace) with bounded long-run mean $\bar{G}^\parallel$, and $G^{s,\perp}$ is zero-mean, has bounded covariance, and is weakly temporally autocorrelated. Then in the steady-state regime ($t \gg \tau_k$) the off-subspace to on-subspace energy ratio in the buffer scales as

$$R(\beta_k) := \frac{\mathbb{E}[\|M_k^{t,\perp}\|_F^2]}{\mathbb{E}[\|M_k^{t,\parallel}\|_F^2]} \asymp \frac{1 - \beta_k}{1 + \beta_k},$$

and $R(\cdot)$ is strictly decreasing on $[0, 1)$. The proportionality constant depends on the noise level and subspace geometry; the formal asymptotic statement is in Appendix C.

A fully rigorous restatement, the $O((1 - \beta_k)/(1 + \beta_k))$ derivation, the matrix-Bernstein upgrade to a high-probability stable-rank envelope, and a discussion of the assumptions are deferred to Appendix C; the singular-value spectra in Figure 2(a) confirm the predicted ordering empirically.

Assumptions and scope. Proposition 1 rests on two assumptions on the gradient stream: (i) a slowly-rotating low-rank parallel subspace $\mathcal{S}$ (the GaLore subspace assumption), and (ii) weak temporal autocorrelation of the perpendicular noise. Figure 2 verifies both directly from saved checkpoints. The singular-value spectra (a) confirm that the narrow-band buffer ($\beta_2=0.998$) decays faster in the high-rank tail than the wide-band buffer ($\beta_1=0.969$). Panels (b,c) show that the top- r subspaces of both buffers remain above the random baseline at lag $T=100$ (overlap 0.445 for β_1 and 0.632 for β_2), while their temporal behaviour differs sharply: the wide-band momentum is already decorrelated ($\cos \approx -0.03$), whereas the

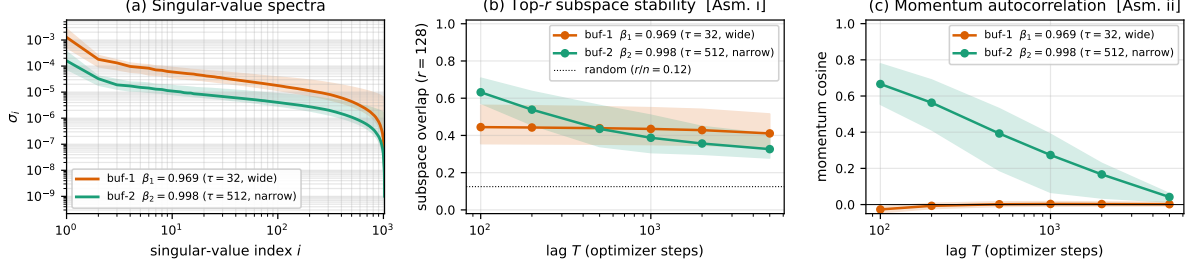


Figure 2: Direct verification of Proposition 1 on saved $K=2$ checkpoints ($\beta_1=0.969$ wide, $\beta_2=0.998$ narrow). **(a)** Per-index singular values of the two momentum buffers at $t=10,172$ from a full-rank $K=2$ Multiscale Muon run (25–75% interquartile range shaded). **(b)** top- r subspace overlap $\frac{1}{r} \|V_k(t)^\top V_k(t+T)\|_F^2$ for each buffer (Assumption i). **(c)** cosine $\langle M_k(t), M_k(t+T) \rangle / (\|M_k(t)\| \|M_k(t+T)\|)$ as a proxy for Assumption ii. Panels (b,c) use the operational S-MuG $K=2$ run (rank-128 narrow-band truncation). Lines are means and bands are IQR over all $(t, t+T)$ reference pairs and parameters; dotted line is the random-subspace baseline $r/n=0.125$.

narrow-band momentum retains substantial cosine (≈ 0.67). This reflects the faster temporal variation of the wide-band signal. The full-rank $K=3$ counterpart shows the same qualitative distinction (Appendix C).

Under GaLore compression, buffer k keeps $\tilde{M}_k^t = \sum_s \beta_k^{t-s} G^{s,\parallel}$ and discards the off-subspace tail; the relative energy error $R(\beta_k)$ is $O(1/\tau_k)$ as $\beta_k \rightarrow 1$ (narrow band: safe) but $\Theta(1)$ as $\beta_k \rightarrow 0$ (wide band: destructive), where $\tau_k := 1/(1 - \beta_k)$. This is the design rule we instantiate in Section 4: *narrow band* $\Leftrightarrow$ *high β* $\Leftrightarrow$ *low effective rank* $\Leftrightarrow$ *safe to compress*.

3.4 Cross-Band Subspace Structure

Proposition 1 and the filter-bank picture above describe the momentum buffers *before* Newton–Schulz orthogonalisation. NS flattens their singular values ($\text{NS}(U\Sigma V^\top) \approx UV^\top$), so spectral magnitudes themselves do not survive into the final update, but it preserves the subspace structure.

For each buffer, let $V_k \in \mathbb{R}^{n \times r}$ denote its top- r singular subspace and decompose $M_k = M_k^{\text{main}} + M_k^{\text{tail}}$, where $M_k^{\text{main}} \equiv M_k V_k V_k^\top$ and $M_k^{\text{tail}} \equiv M_k (I - V_k V_k^\top)$. We analyse the weight matrices in Multiscale-Muon $K=2$ run, using checkpoints every 100 steps from 500 to 10,100. Values are means across parameters; the random subspace baseline is $r/n=0.125$.

The measurements give two structural observations. **(1) Main subspaces overlap across bands:** the steady-state mean $\|V_1^\top V_2\|_F^2/r$ is 0.495 ($4.0\times$ baseline). **(2) Tails are approximately band-specific:** cross-band tail-in-main ratios remain near baseline (0.124 and 0.119), so either band’s main subspace captures little of the other’s residual.

Combining these observations with the filter-bank viewpoint, we make the following conjecture that motivates S-MUG:

Conjecture 1. *The effective information in the narrow-band (i.e. high- β) momentum concentrates in M_2^{main} ; this low-dimensional component corrects the wide-band momentum, while M_2^{tail} is non-informative or harmful.*

4 S-MUG: Selective Multiscale Muon with GaLore

Applying GaLore only to the narrow-band buffer β_2 and retaining the wide-band buffer β_1 at full rank gives the proposed S-MUG.

4.1 Algorithm

Algorithm 1 describes S-MUG for a single weight matrix $W \in \mathbb{R}^{m \times n}$.

Single Newton-Schulz on the combined update. S-MUG forms $s_1 M_1 + s_2 M_2 V^\top$ in the full space and applies a *single* NS pass; the Multiscale Muon baseline keeps per-buffer NS. The combined-NS choice is principled: per-buffer NS would orthogonalise the rank- r surrogate $s_2 M_2 V^\top$ (with $M_2 \in \mathbb{R}^{m \times r}$) in isolation, where NS sees a degenerate spectrum; summing first restores a full-rank input via M_1 . Table 12 (Appendix F) bears this out: with both buffers full-rank the two placements tie (21.71 vs. 21.69), but once the narrow-band buffer is rank-128 truncated, combined-NS overtakes per-buffer NS (21.66 vs. 21.73). The wall-clock gap in Table 3 mixes the added projection/SVD cost with the $K-1$ NS passes saved; Appendix E shows the two effects cancel to within noise at $K=3$. The

Table 1: Pre-NS cross-band subspace trajectory on the operational S-MuG $K=2$ run (means over 96 parameters). Baseline $r/n=0.125$.

step	$\ V_1^\top V_2\ _F^2 / r$	$\ V_2^\top M_1^{\text{tail}}\ _F^2 / \ M_1^{\text{tail}}\ _F^2$	$\ V_1^\top M_2^{\text{tail}}\ _F^2 / \ M_2^{\text{tail}}\ _F^2$
1000	0.391	0.120	0.158
2000	0.459	0.126	0.132
5000	0.498	0.124	0.116
9000	0.510	0.122	0.117
10000	0.533	0.121	0.117
mean ≥ 2000	0.495	0.124	0.119

Algorithm 1 S-MUG ($K=2$): per weight matrix $W \in \mathbb{R}^{m \times n}$. We use the single-NS variant (one NS pass on the combined update).

Require: Learning rate η ; $\beta_1 < \beta_2$; refresh frequency T ; rank r ; NS iterations q

- 1: Initialize $M_1 \in \mathbb{R}^{m \times n} \leftarrow \mathbf{0}$, $M_2 \in \mathbb{R}^{m \times r} \leftarrow \mathbf{0}$;
 $V \in \mathbb{R}^{n \times r}$ deferred until step 1
- 2: $s_k \leftarrow (1 - \beta_k) / \sum_j (1 - \beta_j)$ for $k = 1, 2$
- 3: **for** step $t = 1, 2, \dots$ **do**
- 4: Compute gradient $G^t = \nabla_W \mathcal{L}^t$
- 5: **if** $t = 1$ **or** $t \bmod T = 0$ **then**
- 6: $V_{\text{new}} \leftarrow$ top- r right singular vectors of G^t
- 7: **if** $t \neq 1$ **then**
- 8: $M_2 \leftarrow M_2 (V_{\text{old}}^\top V_{\text{new}})$
- 9: **end if**
- 10: $V \leftarrow V_{\text{new}}$
- 11: **end if**
- 12: $M_1 \leftarrow \beta_1 M_1 + (1 - \beta_1) G^t$; $M_2 \leftarrow \beta_2 M_2 + (1 - \beta_2) (G^t V)$
- 13: $\tilde{U} \leftarrow s_1 M_1 + s_2 (M_2 V^\top)$; $\hat{U} \leftarrow \text{NS}(\tilde{U}, q) \cdot \sqrt{\max(m, n)}$
- 14: $W \leftarrow W - \eta \hat{U}$
- 15: **end for**

$\sqrt{\max(m, n)}$ factor is the standard Muon spectral-norm scaling (Jordan et al., 2024).

Why V is refreshed but M_2 is realigned, not reset. GaLore’s reset-on-refresh would discard $\tau_2=512$ steps of accumulated low-frequency content every $T=200$ steps. We instead realign via $M_2 \leftarrow M_2 (V_{\text{old}}^\top V_{\text{new}})$ (Algorithm 1, line 8), which is dimensionally consistent, neutralises SVD sign ambiguity, and (since $T < \tau_2$ keeps $V_{\text{old}}^\top V_{\text{new}}$ close to I) preserves the accumulated content under the low-rank-subspace-stability assumption GaLore relies on (Proposition 1). Refresh sensitivity over T is reported in Table 8.

Table 2: Optimizer momentum state per weight matrix, $K=2$, $r/n=0.125$. Counts both EMA buffers *and* the persistent GaLore basis V . The $1.250\times$ ratio assumes square $m=n$; for general $m \times n$ the ratio is $1+r/n+r/m$ (all 201M weight matrices studied here are square, $m=n=1024$).

Optimizer	State	Rel.
Muon (single β)	mn	$1.000\times$
Multiscale Muon ($K=2$)	$2mn$	$2.000\times$
S-MUG ($K=2$)	$mn + mr + nr$	$1.250\times$

4.2 Memory Analysis

Table 2 compares optimizer-state memory per weight matrix $W \in \mathbb{R}^{m \times n}$ for $K=2$, rank ratio $r/n = 0.125$. We count *all* persistent optimizer state, including the GaLore basis $V \in \mathbb{R}^{n \times r}$ that is read every step and refreshed every T steps.

S-MUG reduces the $2\times$ memory overhead of Multiscale Muon to $1.25\times$ relative to single-scale Muon, recovering 75% of the added overhead at no average quality cost (Section 5.3).

5 Experiments

We organize our experiments around three core questions: (Q1) does multi-band coverage outperform any single-band filter; (Q2) what quality-memory trade-off does selective compression provide, and does it transfer with model scale; (Q3) why is selective compression safe, and which frequency band can be compressed. We additionally test two robustness properties: whether the rule composes with per-buffer Muon improvements, and how sensitive it is to the projection-refresh interval.

5.1 Setup

We pretrain a 201 M-parameter GPT-2-style transformer (16 layers, 8 heads, $d_{\text{model}}=1024$, context 1024) on FineWeb (BPE vocab 50 304) via nanoGPT (Karpathy, 2022). Cosine LR

(peak $\eta=0.0018$, warmup 500, warmdown 5812 steps), batch $480 \times 1024 \approx 492$ K tokens/step, 5B tokens total (10172 steps); $q=5$ Newton-Schulz iterations for all Muon variants. Multiscale Muon: $K=2$, $(\beta_1, \beta_2)=(0.96875, 0.99805)$ ($\tau_1=32$, $\tau_2=512$ steps), scales $s_k \propto (1-\beta_k)$. GaLore: rank ratio $r/n=128/1024=0.125$ on all linear weight matrices, projection refreshed every $T=200$ steps; embeddings always use AdamW. Baselines: AdamW ($\beta=(0.9, 0.95)$), single- β Muon ($\beta=0.98$), full-rank Multiscale Muon, and S-MUG (full-rank β_1 , GaLore $r=128$ on β_2). All reported numbers and shaded bands use mean $\pm$ standard deviation across $n=3$ seeds per configuration (seeds {42, 43, 44}). Hardware, software versions, and the complete hyperparameter table are listed in Appendix A.

5.2 Q1: Does Multi-Band Coverage Outperform Any Single-Band Filter?

Figure 3 shows validation perplexity over training for Muon, Multiscale Muon ($K \in \{2, 3\}$) and S-MUG ($K \in \{2, 3\}$).

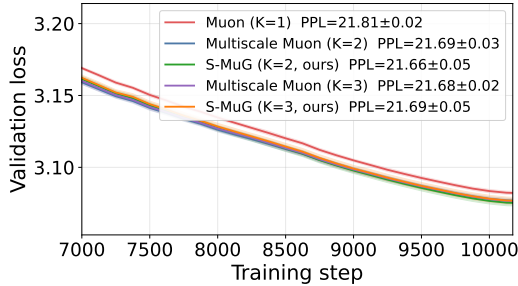


Figure 3: Training curves (mean $\pm$ s.d. across $n=3$ seeds, shaded band): Muon $K=1$, Multiscale Muon ($K \in \{2, 3\}$), and S-MUG ($K \in \{2, 3\}$). Final PPL values are reported in Table 3; S-MUG $K=2$ ties the best multiscale variant within 1 s.d. while using only $1.25 \times$ Muon’s optimizer memory.

Going from $K=1$ (PPL = 21.81 ± 0.02) to $K=2$ (PPL = 21.69 ± 0.03) yields a gain of ≈ 0.12 PPL points, and no single- β Muon variant matches $K=2$ regardless of β choice. This confirms that the gradient sequence carries meaningful energy at both short and long timescales, only jointly exploitable by a multi-band filter bank. The $K=1$ – 5 sweep in Appendix F (Table 13) shows that the gain saturates at $K=2/3$.

Effect of (β_1, β_2) and the mixing coefficient λ . The filter-bank perspective predicts an interior optimum for the narrow-band weight λ (parameterised

Table 3: Validation perplexity (mean $\pm$ s.d. over $n=3$ seeds) and peak optimizer state, 200M GPT-2 pre-training (5B tokens). *Rel. Mem.*: analytical per-rank optimizer-state ratio rel. to single Muon. *Peak Mem.*: measured per-GPU memory (Muon variants shard momentum, ZeRO-1 style). *Wall*: end-to-end training time in hours on $8 \times \text{MI300X}$ (per-step average from each run’s full timing log). **Bold** numbers mark the best entry in each column; the proposed S-MUG rows are bolded in the leftmost column for emphasis.

Optimizer	PPL	Rel.	Peak	Wall
AdamW	22.70 ± 0.06	$2.00 \times$	12.79	3.33
Muon ($K=1$)	21.81 ± 0.02	$1.00 \times$	10.31	3.26
Muon+GaLore ($K=1$, $r=128$)	29.43 ± 0.02	$0.25 \times$	10.39	3.52
MS Muon ($K=2$)	21.69 ± 0.03	$2.00 \times$	10.49	3.28
MS Muon ($K=3$)	21.68 ± 0.02	$3.00 \times$	10.68	3.29
S-MUG $K=2$	21.66 ± 0.05	$1.25 \times$	10.42	3.32
S-MUG $K=3$	21.69 ± 0.05	$2.25 \times$	10.61	3.29

as $s_1 = 1 - \lambda$, $s_2 = \lambda$). To probe this we fit a piecewise-constant spectral score $g(\omega)$ ($N_{\text{bands}}=5$) jointly against the $K=1$ single-buffer sweep and the $K=2$ (τ_1, τ_2, λ) table; the recovered λ^* predictions correlate with observed optima across (τ_1, τ_2) pairs (Spearman $\rho=0.97$). We treat this fit as a *post-hoc diagnostic* of the filter-bank picture, not as a closed-form recipe: the spectral score is fit to the same family of (τ_1, τ_2) runs that it is then evaluated on, and Appendix H proves no non-negative linear spectral score on equal-energy normalised filters can reproduce the interior optimum in closed form (in part because of the per-buffer NS non-linearity). Plots and methodology are deferred to Appendix G.

Takeaway. A two-band filter bank ($K=2$) consistently outperforms any single-bandwidth optimizer because the gradient sequence contains meaningful energy at both short-horizon and long-horizon timescales simultaneously.

5.3 Q2: What Quality–Memory Trade-off Does Selective Compression Provide?

S-MUG ($K=2$) reaches PPL 21.66 ± 0.05 , statistically tied with full-rank Multiscale Muon $K=2$ (21.69 ± 0.03), while using $1.25 \times$ the memory of single Muon vs. $2.00 \times$ for full Multiscale Muon. This is a $4 \times$ reduction in the added memory overhead at no average quality cost. Both S-MUG and Multiscale Muon substantially outperform single-scale baselines (AdamW: 22.70 ± 0.06 ; Muon $K=1$: 21.81 ± 0.02). The single-scale Muon+GaLore baseline ($K=1$, same rank $r=128$ as S-MUG’s narrow-band branch) is much worse

Table 4: Validation perplexity at 756 M parameters, 10,172 steps, mean $\pm$ s.d. across $n=2$ seeds (42, 43). Wall in hours on 8 $\times$ MI300X (single-seed wall-time).

Optimizer	Val. loss $\downarrow$	Val. PPL $\downarrow$	Wall (h)
Muon ($K=1$)	2.911 $\pm$ 0.000	18.38 $\pm$ 0.01	5.87
Multiscale Muon ($K=2$)	2.912 $\pm$ 0.000	18.40 $\pm$ 0.01	5.99
S-MuG ($K=2$)	2.910$\pm$0.001	18.36$\pm$0.01	5.95

at 29.43 $\pm$ 0.02 PPL, confirming that low-rank projection on a wide-band buffer is destructive (Section 5.6) and that the multiscale filter bank is a prerequisite for safely compressing optimizer state. S-MuG pays only a $\approx 1.4\%$ wall-time overhead over the full-rank Multiscale Muon baseline at $K=2$ (3.32 h vs. 3.28 h) for the per-step rank- r projection and periodic truncated-SVD refresh, and is essentially free at $K=3$ (3.29 h vs. 3.29 h) once the projection cost is amortised against K Newton-Schulz iterations; the GaLore-on-both-buffers baseline (Table 7, last block) is considerably worse on PPL. *Rel. Mem.* is an analytical optimizer-state ratio (per-parameter); measured peak GPU memory is similar across rows (10.31/10.42/10.49 GB) because activations and sharded weights dominate under 8-way ZeRO-1 at this scale, so the headline benefit at this scale is the optimizer-state ratio and the resulting headroom for larger microbatch or depth.

5.4 Scaling validation at 756 M parameters

To check that the ordering observed at 201 M (Table 3) is not specific to that scale, we rerun the headline configurations on a 756 M-parameter variant of the same GPT-2 architecture (16 heads, $d_{\text{model}}=1536$, depth 24; identical tokenizer, data, schedule, and step budget). Per-configuration cost is $\sim 3\times$ that of the 201 M runs, so we report $n=2$ seeds (42, 43) per headline configuration.

At 756 M, the three configurations are within seed-level noise on this $n=2$ sample (PPL 18.36–18.40, all standard deviations ≤ 0.01), so the table should be read as scaling *consistency* rather than a statistically significant ordering. With that caveat: full-rank Multiscale Muon *does not improve over single Muon* (PPL 18.40 $\pm$ 0.01 vs. 18.38 $\pm$ 0.01); S-MuG ($K=2$) is the only configuration to (slightly) clear both baselines on per-seed average (PPL 18.36 $\pm$ 0.01), at 1.25 $\times$ rather than 2.00 $\times$ memory. The selective-compression rule of Section 4 appears to do double duty at this scale: it cuts memory while preserving the multi-band gain that the

unselective baseline fails to deliver. At the larger 1 B scale (Table 5), the picture sharpens: full-rank MS Muon (WikiText PPL 30.01 $\pm$ 0.03) does outperform single Muon (30.10 $\pm$ 0.00), confirming that the multi-band advantage emerges with scale, and S-MuG further improves to 29.87 $\pm$ 0.01 — demonstrating the selective-compression rule provides consistent benefit at every scale tested.

5.5 Scaling and downstream transfer

To confirm that the scale-up behaviour transfers downstream, we plug the headline optimizers into separate 1 B and 3 B pretraining pipelines and evaluate them with the same downstream suite. At 1 B, we evaluate the three headline optimizers at iteration 10 000 on $n=2$ seeds (42, 43; per-step token budget $4096 \times 1024 \approx 4.19$ M, total tokens consumed ≈ 42 B per seed), reporting $K=2$ only at this scale.

We treat this section as *exploratory transfer evidence on $n=2$ seeds*: LAMBADA itself shows large seed scatter (e.g. MS Muon LAMBADA-PPL std ± 1.0 and Muon LAMBADA-PPL flipping leader across the two seeds), so we restrict robust claims to the lower-variance metrics. On the WikiText and in-domain LM-loss columns the ordering $\text{S-MuG} \leq \text{MS Muon} \leq \text{single Muon}$ holds in *both* seeds individually, and S-MuG is best on 5/6 LM columns overall (Table 5). Downstream zero-/few-shot evaluation (Appendix D, Table 10) is consistent: Muon’s Code average (7.9 $\pm$ 1.1%) is clearly below both multiscale variants (9.3 $\pm$ 0.7% MS Muon, 9.2 $\pm$ 0.8% S-MuG), Reasoning is tied within noise across all three (45.1–45.5%), and overall benchmark average rises monotonically from Muon (36.9 $\pm$ 0.4%) to MS Muon (37.1 $\pm$ 0.4%) to S-MuG (37.5 $\pm$ 0.7%), again in both seeds individually. These observations confirm that selectively compressing the narrow-band buffer preserves the multiscale gain on the low-variance LM metrics and slightly improves the aggregate downstream average; per-task swings remain noisy, motivating larger seed counts in future work.

3B transfer. For the 3 B scale-transfer evaluation, we evaluate at iteration 30,000 using the same language-modelling metrics. Because of the computational cost of this model size, we run only single-buffer Muon and S-MuG, with one seed per optimizer; these results are therefore directional rather than a multi-seed estimate. S-MuG signif-

Table 5: 1 B GQA language-modelling metrics at iter 10 000 (mean $\pm$ std over $n=2$ seeds, $\{42, 43\}$). **Bold** = best per column. ID = in-domain validation.

Optimizer	WikiText PPL $\downarrow$	WikiText bpb $\downarrow$	LAMBADA PPL $\downarrow$	LAMBADA acc $\uparrow$	ID PPL $\downarrow$	ID NLL $\downarrow$
Muon ($K=1$)	30.10 $\pm$ 0.00	0.919 $\pm$ 0.000	37.44 $\pm$ 0.30	0.342$\pm$0.004	3.399 $\pm$ 0.001	1.224 $\pm$ 0.000
MS Muon ($K=2$)	30.01 $\pm$ 0.03	0.918 $\pm$ 0.000	37.53 $\pm$ 1.02	0.338 $\pm$ 0.004	3.395 $\pm$ 0.001	1.222 $\pm$ 0.000
S-MuG ($K=2$)	29.87$\pm$0.01	0.916$\pm$0.000	36.60$\pm$1.46	0.339 $\pm$ 0.006	3.387$\pm$0.000	1.220$\pm$0.000

icantly improves WikiText PPL (22.24 vs. 24.01), LAMBADA PPL (17.70 vs. 22.09), and in-domain evaluation (3.00 vs. 3.11 PPL), confirming that the multiscale advantage carries over to larger parameters (Table 6). The detailed Code and Reasoning results are reported in Appendix D.

Q2 Takeaway. Across 201 M, 756 M, 1 B, and the 3 B transfer check, selective compression preserves or improves the quality of the multiscale recipe at substantially lower optimizer-state cost.

5.6 Q3: Why Is Selective Compression Safe?

Compression ablation. Table 7 directly tests which $K=2$ buffer can be compressed. Applying GaLore to the wide-band β_1 buffer at any rank consistently increases perplexity, confirming that compressing a high-rank, wide-band state is destructive (PPL rises to 23.43 at $r_1=128$). Applying GaLore to the narrow-band β_2 buffer provides meaningful memory savings at minimal quality loss: a $r/n \in \{0.125, 0.25, 0.5, 1.0\}$ sweep (compressing β_2 only, $K=2$) gives PPL $\{21.66, 21.73, 21.78, 21.71\}$ respectively ($r/n=0.125$ is the multi-seed smug mean from Table 3; others are single-seed), all within 0.08 PPL of the full-rank baseline, so the default $r/n=0.125$ offers the best memory reduction at indistinguishable quality.

Selectivity vs. projecting both buffers. Routing the wide-band (β_1) buffer through GaLore actively hurts quality *as soon as* it is actually compressed. At matched $1.25\times$ memory, symmetric ($r_1=r_2=320$) and asymmetric ($r_1=512, r_2=128$) routings reach PPL 25.13 (+3.47 over S-MuG) and 23.42 (+1.76), both far worse than S-MuG’s 21.66 ± 0.05 (Table 3). The boundary case $r_1=1024=n$ is a *no-op* compression on β_1 (since $V_1 V_1^\top = I$ at full rank, so the projection becomes the identity) and recovers 21.67 at $2.25\times$ memory; this rules out the projection *mechanism* as the source of harm and isolates it to genuine wide-band compression. The last block of Table 7 then shows

the full monotone gradient $21.67 \rightarrow 23.42 \rightarrow 29.41$ as r_1 shrinks $1024 \rightarrow 512 \rightarrow 128$, confirming that further wide-band compression widens the gap monotonically rather than saturating.

Takeaway. The fast/wide-band buffer (β_1) must remain at full rank; the slow/narrow-band buffer (β_2) is safely compressible with GaLore at $r/n=0.125$.

5.7 Robustness: Refresh Sensitivity and Composability

As a supplementary robustness study, we first test the projection-refresh interval. The basis is refreshed every T steps; all five settings in Table 8 lie within 0.01 PPL, so $T=200$ is a practical default rather than a narrowly tuned optimum.

We then test whether two distinct improvements can be combined. NorMuon (Li et al., 2025) and TEON (Zhang et al., 2026) modify the update rule of a single-buffer Muon optimizer, whereas S-MuG selectively compresses the narrow-band buffer of a multi-buffer optimizer. At 201 M, applying the same selective assignment (full-rank wide band; rank-128 narrow band) improves NorMuon from 21.78 to 21.65 PPL and TEON from 21.87 to 21.72 PPL over three seeds (Table 9). This supports a composable, vertical relationship rather than treating these methods as competing alternatives.

6 Related Work

Spectral and orthogonal optimizers. Muon (Jordan et al., 2024) approximates an orthogonal-gradient update via Newton-Schulz under the spectral-norm trust-region view of Bernstein and Newhouse (2024), with scaling behaviour studied in Liu et al. (2025); related preconditioned methods include Shampoo/SOAP (Gupta et al., 2018; Vyas et al., 2024). Multiscale Muon (Pappone and Selvaraj, 2025) extends Muon with K momentum buffers at distinct decay rates and is the immediate point of departure for our work.

Table 6: 3 B GQA language-modelling metrics at iteration 30,000 (one seed per optimizer). ID = in-domain validation.

Optimizer	WikiText PPL	WikiText bpb	LAMBADA PPL	LAMBADA acc.	ID PPL↓	ID NLL↓
Muon	24.0139	0.8576	22.0885	0.4058	3.1100	1.1346
S-MuG	22.2375	0.8368	17.6961	0.4283	2.9989	1.0982

Table 7: $K=2$ compression ablation, validation PPL. “ $\times$ ” = full-rank EMA; numbers indicate GaLore rank applied to that buffer. **Bold** marks the proposed S-MUG configuration. Memory column counts both EMA buffers and the GaLore basis V (square $m=n=1024$).

β_1	β_2	r_1	r_2	PPL	Mem.
$\times$	$\times$	—	—	21.69	2.00 $\times$
$\times$	$\checkmark$	—	128	21.66	1.25$\times$
$\times$	$\checkmark$	—	256	21.73	1.50 $\times$
$\times$	$\checkmark$	—	512	21.78	2.00 $\times$
$\checkmark$	$\times$	128	—	23.43	1.25 $\times$
$\checkmark$	$\times$	256	—	22.77	1.50 $\times$
<i>Both buffers projected through GaLore (combined NS):</i>					
$\checkmark$	$\checkmark$	1024	128	21.67	2.25 $\times$
$\checkmark$	$\checkmark$	512	128	23.42	1.25 $\times$
$\checkmark$	$\checkmark$	320	320	25.13	1.25 $\times$
$\checkmark$	$\checkmark$	128	128	29.41	0.50 $\times$

Table 8: Projection-refresh sensitivity for S-MuG $K=2$ at 201 M (single seed).

T	50	100	200	400	800
PPL	21.66	21.66	21.66	21.67	21.66

Low-rank optimizer state and adapters. GaLore (Zhao et al., 2024) compresses the optimizer state by projecting gradients into a slowly-rotating top- r subspace; Flora (Hao et al., 2024), COAP (Xiao et al., 2025), and MLorc (Shen et al., 2026) are alternative low-rank state-compression mechanisms (random projections, correlation-aware, and direct momentum compression respectively), and Pan et al. (2025) analyse and correct the bias of low-rank gradient projection. ReLoRA/LoRA (Lialin et al., 2024; Hu et al., 2022) instead decompose weights. S-MUG is orthogonal: rather than modifying *how* a single buffer is compressed, it asks *which* of K multi-scale buffers should be compressed, and gives a frequency-domain answer (Section 3.3); the rule could be paired with any of the above mechanisms.

Frequency-domain analyses of momentum. Momentum-as-low-pass-filter is implicit in classical analyses (Nesterov, 1983; Sutskever et al., 2013) and modern bandwidth/plasticity studies (Lyle et al., 2024); to our knowledge, the multi-

Table 9: Composition with single-buffer Muon variants at 201 M ($n=3$ seeds). S-MUG supplies selective multi-buffer compression; NorMuon and TEON supply per-buffer update improvements.

Optimizer	Val. PPL ↓
NorMuon	21.78 $\pm$ 0.03
NorMuon + S-MUG	21.65$\pm$0.01
TEON	21.87 $\pm$ 0.02
TEON + S-MUG	21.72$\pm$0.05

β subband filter-bank synthesis (Vaidyanathan, 1993) and its low-rank consequence (Proposition 1) are new and underlie S-MUG.

7 Conclusion

We introduced S-MUG, a selective momentum compression framework grounded in a frequency-domain view: Multiscale Muon acts as a subband filter bank over the gradient stream whose per-channel compressibility follows directly from its bandwidth (Proposition 1). Specifically, narrow-band (high- β) channels collect the slow, low-frequency, and low-rank components of gradient history and are safely compressible by GaLore. In contrast, wide-band (low- β) channels track high-frequency, near-full-rank transient directions and must remain uncompressed to safeguard convergence.

Empirically, S-MUG ($K=2$) reduces Multiscale Muon’s added momentum overhead by $4\times$ ($1.25\times$ vs. $2.00\times$ single-Muon memory) at no average quality loss. This quality-memory preservation scales consistently from 201 M up to 756 M dense, as well as 1 B and 3 B GQA pretraining. Furthermore, we demonstrated that S-MUG’s selection rule is robust to basis refresh rates and is highly composable, bringing substantial vertical improvements when paired with single-buffer orthogonal optimizers like NorMuon and TEON. Beyond its immediate practical utility in lowering the hardware requirements of orthogonal training, we conjecture that the bandwidth–rank selectivity principle represents a general design paradigm that extends to all multi-buffer EMA optimizers.

Limitations

Scale, seeds, and the nature of the trade-off.

Multi-seed evidence is at 201 M / 5 B tokens ($n=3$ seeds per configuration), at 756 M for the three headline configurations ($n=2$, seeds 42 and 43), and at 1 B GQA for the three headline configurations ($n=2$, seeds 42 and 43). Behavior beyond 1 B, on MoE / encoder-decoder, or non-language modalities is unverified, and Proposition 1 assumes a slowly-rotating low-rank gradient subspace that may weaken under sharp distribution shifts. The contribution is asymmetric: relative to single-scale Muon, S-MUG *adds* optimizer memory ($1.25\times$ vs. $1.00\times$) and the payoff is the multi-band PPL gain; relative to full-rank Multiscale Muon, S-MUG preserves that gain while cutting the added overhead $4\times$. Absolute PPL gaps are small (≤ 0.04 at 756 M; aggregate downstream gains at 1 B are ≈ 0.6 pt overall average), so the contribution is best characterised as a favourable quality–memory operating point rather than a strict PPL win.

Wall-clock, systems, hyperparameters. S-MUG pays only $\approx 1.4\%$ wall-time over full-rank Multiscale Muon at $K=2$ (Table 3); the freed memory enables a larger microbatch but not a multiplicative throughput gain. Interaction with FP8/INT8 state quantisation (Dettmers et al., 2022), offload, and pipeline parallelism is unverified. The mixing λ , decays (β_1, β_2) , rank ratio r/n , and refresh interval T remain empirical: our spectral score ranks λ^* across (τ_1, τ_2) (Spearman $\rho=0.97$), but Appendix H proves no non-negative linear spectral score predicts the interior optimum in closed form, so expect a small grid search at a new scale.

Cross-mechanism generality. The selection rule (*compress only the narrow-band buffer*) is presented as the conceptual contribution, but we instantiate it only with GaLore and only with Newton-Schulz post-processing; we do not benchmark the rule on alternative low-rank/state-compression mechanisms (e.g. COAP (Xiao et al., 2025), MLorc (Shen et al., 2026), or 8-bit state quantisation (Dettmers et al., 2022)), nor against the unbiased low-rank correction of Pan et al. (2025). Establishing that the selection rule transfers across projection mechanisms is left to future work. The headline comparison in Table 3 also changes NS placement (single combined NS in S-MUG vs. per-buffer NS in full-rank Multiscale Muon); as argued in §4 this is a principled consequence of the

low-rank GaLore branch (per-buffer NS would orthogonalise a degenerate, rank- r surrogate), and the full 2×2 design in Appendix F (Table 12) confirms the four cells span only 0.07 PPL with the small interaction favouring combined-NS once the narrow-band buffer is rank-truncated.

Ethics Statement

S-MUG is a drop-in replacement for an existing optimizer (Multiscale Muon) on standard language-model pretraining; it introduces no new training data, model architecture, or deployment pathway, and therefore does not alter the downstream risk profile of the LMs to which it is applied (toxic generation, factual error, bias, dual-use). Its primary practical effect is reducing per-device optimizer-state memory by $\approx 38\%$ relative to full-rank Multiscale Muon at $K=2$ ($1.25\times$ vs. $2.00\times$ single-Muon memory), which can lower the compute barrier for academic and small-industry groups to reproduce multi-scale-momentum results that would otherwise require a larger memory budget. We see no specific harms that are unique to S-MUG beyond those already documented for LM pretraining at this scale. All 201 M and 756 M experiments use the public FineWeb corpus; the 1 B GQA transfer test uses an in-house pretraining corpus subject to its own data-governance policies. Experiments use a single-node $8\times$ accelerator setup, well within standard pretraining energy budgets reported by prior 200 M–1 B-parameter optimizer studies.

Acknowledgements

We sincerely thank our colleagues and mentors at Microsoft Research Asia for their invaluable guidance, feedback, and encouragement throughout this project. We are also grateful to Microsoft Research Asia for providing the substantial computational resources required to run the large-scale pretraining models and downstream evaluation pipelines.

References

- Jeremy Bernstein and Laker Newhouse. 2024. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *International Conference on Learning Representations (ICLR)*.
- Vineet Gupta, Tomer Koren, and Yoram Singer. 2018. Shampoo: Preconditioned stochastic tensor optimization.

- tion. In *International Conference on Machine Learning (ICML)*.
- Yongchang Hao, Yanshuai Cao, and Lili Mou. 2024. Flora: Low-rank adapters are secretly gradient compressors. In *International Conference on Machine Learning (ICML)*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. 2024. Muon: An optimizer for hidden layers in neural networks. <https://kellerjordan.github.io/posts/muon/>. Code: <https://github.com/KellerJordan/Muon>.
- Andrej Karpathy. 2022. nanogpt. <https://github.com/karpathy/nanoGPT>.
- Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. 2025. NorMuon: Making Muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*.
- Vladislav Lialin, Sherin Muckatira, Namrata Deshpande, and Anna Rumshisky. 2024. ReLoRA: High-rank training through low-rank updates. In *International Conference on Learning Representations (ICLR)*.
- Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, and 9 others. 2025. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*.
- Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. 2024. Disentangling the causes of plasticity loss in neural networks. *arXiv preprint arXiv:2402.18762*.
- Yurii Nesterov. 1983. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547.
- Alan V. Oppenheim and Ronald W. Schaffer. 1999. *Discrete-Time Signal Processing*, 2nd edition. Prentice Hall.
- Rui Pan, Yang Luo, Yuxing Liu, Yang You, and Tong Zhang. 2025. Unbiased gradient low-rank projection. *arXiv preprint arXiv:2510.17802*.
- Francesco Pappone and Devan Selvaraj. 2025. Multiscale Muon. Blog post, <https://publish.obsidian.md/ueaj/Machine+Learning/Pretraining/Multiscale+Muon>.
- Wei Shen, Zhang Yaxiang, Minhui Huang, Mengfan Xu, Jiawei Zhang, and Cong Shen. 2026. MLorc: Momentum low-rank compression for memory efficient large language model adaptation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning (ICML)*.
- Joel A. Tropp. 2015. *An Introduction to Matrix Concentration Inequalities*, volume 8 of *Foundations and Trends in Machine Learning*. Now Publishers.
- P. P. Vaidyanathan. 1993. *Multirate Systems and Filter Banks*. Prentice Hall.
- Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. 2024. SOAP: Improving and stabilizing shampoo using Adam. *arXiv preprint arXiv:2409.11321*.
- Jinqi Xiao, Shen Sang, Tiancheng Zhi, Jing Liu, Qing Yan, Yuqian Zhang, Linjie Luo, and Bo Yuan. 2025. COAP: Memory-efficient training with correlation-aware gradient projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, Dongyang Li, Yupeng Su, Sijia Liu, and Zheng Zhang. 2026. TEON: Tensorized orthonormalization beyond layer-wise Muon for large language model pre-training. *arXiv preprint arXiv:2601.23261*.
- Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning (ICML)*. Oral presentation.

A Full Experimental Details

All runs use the nanoGPT codebase (Karpathy, 2022) on a single 8×MI300X-192GB node. **201 M GPT-2**: 16 layers, 8 heads, $d_{\text{model}}=1024$, context (block size) 1024, BPE vocabulary size 50 304, 201 M trainable parameters. **756 M GPT-2**: 24 layers, 16 heads, $d_{\text{model}}=1536$, context (block size) 1024, BPE vocabulary size 50 304, 756 M trainable parameters. **1 B GQA**: LLaMA-style architecture with 16 layers, 16 query attention heads, 4 key-value heads (grouped-query attention ratio 4:1), $d_{\text{model}}=2048$, head dimension 128, vocabulary size 200 064, 1.2 B trainable parameters. **3 B GQA**: LLaMA-style architecture with 28 layers, 24 query attention heads, 8 key-value heads (grouped-query

attention ratio 3:1), $d_{\text{model}}=3072$, head dimension 128, vocabulary size 200 064, 3.2 B trainable parameters.

Schedule: cosine LR with peak $\eta = 0.0018$, warmup 500 steps, warmdown 5 812 steps, total 10 172 steps at ≈ 492 K tokens/step (5 B tokens). Muon variants use $q=5$ Newton-Schulz iterations. Multiscale Muon: $K=2$, $(\beta_1, \beta_2) = (0.96875, 0.99805)$ corresponding to EMA time constants $(\tau_1, \tau_2) = (32, 512)$, $s_k \propto (1 - \beta_k)$ then renormalized so $\sum_k s_k = 1$. GaLore: rank ratio $r/n = 0.125$, V refreshed every $T=200$ steps via truncated SVD on G^t . Embeddings always use AdamW. For multi-seed entries we use seeds $\{42, 43, 44\}$. Common settings: weight decay 0.1 on linear weights only, no dropout, GPT-2 default init, bfloat16 mixed-precision.

Licenses and intended use. All artifacts we use are publicly released for research. The FineWeb corpus is distributed under the Open Data Commons Attribution License (ODC-By 1.0); the nanoGPT codebase (Karpathy, 2022) and the Muon reference implementation are released under the MIT license; GaLore (Zhao et al., 2024) is released under Apache-2.0; the Multiscale Muon (Pappone and Selvaraj, 2025) reference is released under a permissive open-source license. Our use, namely academic language-model pretraining and optimizer benchmarking on public web data, is consistent with the stated terms of all of these artifacts. We do not curate or release any new dataset: for the FineWeb experiments, PII anonymisation (email and public-IP regex replacement) and NSFW/toxic-content filtering are inherited from FineWeb’s published curation pipeline. The 1 B GQA experiments (Section 5.5) use a separate in-house pretraining corpus subject to its own data-governance policies (PII filtering, NSFW removal); we do not redistribute it and use it only as a downstream-transfer test of the optimizer rather than as a contribution of new data.

B Newton-Schulz for Non-Square Matrices

The quintic Newton-Schulz iteration with Muon scalars $(a, b, c) = (3.4375, -4.375, 2.1875)$ (Jordan et al., 2024) is a fixed-step orthogonalization heuristic that drives singular values toward 1 (the iteration produces near-unitary outputs but does not converge to the exact polar factor $UV^\top$ in the strict sense). For tall-thin GaLore-compressed buffers

$X \in \mathbb{R}^{m \times r}$ ($m \gg r$) we always evaluate it in the right-grouped form

$$X \leftarrow X(aI_r + bX^\top X + c(X^\top X)^2),$$

which never instantiates the $m \times m$ matrix $XX^\top$ and instead works with $X^\top X \in \mathbb{R}^{r \times r}$ at cost $O(mr^2 + r^3)$ per iteration. Convergence properties are inherited from the square case because the iteration only acts on $X^\top X$. Reconstructing the full-space update is then $\tilde{U}V^\top \in \mathbb{R}^{m \times n}$.

C Singular-Value Spectra and Proof of Proposition 1

Figure 2 (main paper, Section 3.3) plots the singular-value spectra; we restate the methodology and proof here.

Stable rank. We use the standard stable-rank surrogate $\text{srank}(M) = \|M\|_F^2 / \|M\|_2^2$, which lower-bounds the numerical rank, is invariant under orthogonal transforms, and is dimensionless.

Setting and assumptions. Fix a layer with weight $W \in \mathbb{R}^{m \times n}$ and let $G^s \in \mathbb{R}^{m \times n}$ be the per-step stochastic gradient, $s = 1, \dots, t$. Write $\tau_k = 1/(1 - \beta_k)$. The per-buffer EMA with momentum $\beta_k \in [0, 1)$ produces

$$M_k^t = (1 - \beta_k) \sum_{s=1}^t \beta_k^{t-s} G^s.$$

We write the orthogonal decomposition $G^s = \bar{G}^\parallel + \xi_\parallel^s + G^{s,\perp}$, where $\bar{G}^\parallel = \mathbb{E}[\Pi_{\mathcal{S}} G^s]$ is the (slowly drifting) deterministic component lying in a rank- r subspace $\mathcal{S} \subseteq \mathbb{R}^{m \times n}$, $\xi_\parallel^s = \Pi_{\mathcal{S}} G^s - \bar{G}^\parallel$ is the zero-mean fluctuation inside $\mathcal{S}$, and $G^{s,\perp} = (I - \Pi_{\mathcal{S}})G^s$ is the perpendicular fluctuation. We assume:

- (A1) *Mean lies in the subspace:* $\mathbb{E}[G^s] \in \mathcal{S}$, equivalently $\mathbb{E}[G^{s,\perp}] = 0$ for every s . This identifies $\bar{G}^\parallel$ with the deterministic mean of the gradient and makes the parallel/perpendicular decomposition orthogonal in L^2 .
- (A2) *Slow rotation of $\mathcal{S}$:* $\mathcal{S}$ varies on a timescale $T_{\mathcal{S}} \gg \tau_k$, so $\bar{G}^\parallel$ may be treated as time-stationary on a window of length $\Theta(\tau_k)$. Equivalently, $\|\Pi_{\mathcal{S}}^t - \Pi_{\mathcal{S}}^{t-\tau_k}\|_2 = o(1)$, where $\Pi_{\mathcal{S}}^t$ is the orthogonal projector onto $\mathcal{S}$ at time t .
- (A3) *Bounded second moments:* the in-subspace fluctuation is bounded above by a constant $\sigma_\parallel^2$ in spectral norm of $\mathbb{E}[\xi_\parallel^s (\xi_\parallel^s)^\top]$,

and the perpendicular component satisfies the trace bound $\mathbb{E}[\|G^{s,\perp}\|_F^2] \leq n\sigma_\perp^2$ together with the two-sided spectral-norm bounds $\mathbb{E}[(G^{s,\perp})^\top G^{s,\perp}] \preceq \sigma_\perp^2 I_n$ and $\mathbb{E}[G^{s,\perp}(G^{s,\perp})^\top] \preceq \sigma_\perp^2 I_m$, uniformly in s . The latter pair (column- and row-variance) is what matrix Bernstein needs; the trace bound is what the energy-ratio computation needs.

(A4) Weak temporal autocorrelation: $\sum_{u \geq 0} |\rho(u)| \leq C_\rho < \infty$, where $\rho(u)$ is the lag- u scalar autocorrelation of the noise components, normalised so $|\rho(0)| = 1$. This rules out persistent low-frequency noise but allows finite-range correlation.

Formal restatement of Proposition 1. Under assumptions (A1)–(A4) and the EMA update $M_k^t = (1 - \beta_k) \sum_{s=1}^t \beta_k^{t-s} G^s$ with $\beta_k \in [0, 1)$, write $M_k^t = M_k^{t,\parallel} + M_k^{t,\perp}$ for its orthogonal decomposition with respect to the rank- r subspace $\mathcal{S}$. Then in the steady-state regime $t \gg \tau_k = 1/(1 - \beta_k)$:

(i) *Energy ratio.* The off-subspace to on-subspace energy ratio satisfies

$$R(\beta_k) \triangleq \frac{\mathbb{E}[\|M_k^{t,\perp}\|_F^2]}{\mathbb{E}[\|M_k^{t,\parallel}\|_F^2]} = \kappa \frac{1-\beta_k}{1+\beta_k} (1+o(1)),$$

with $\kappa = n C_\rho \sigma_\perp^2 / \|\bar{G}^\parallel\|_F^2$, and $R(\cdot)$ is strictly decreasing on $[0, 1)$ with $R(\beta_k) \rightarrow 0$ as $\beta_k \rightarrow 1^-$.

(ii) *Stable-rank envelope.* Additionally assuming a.s. boundedness $\|G^{s,\perp}\|_2 \leq L$ (e.g. gradient clipping), for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\text{srank}(M_k^t) \leq \text{srank}(\bar{G}^\parallel) + o(1) \leq r,$$

in the limit $\beta_k \rightarrow 1^-$, where $\text{srank}(M) = \|M\|_F^2 / \|M\|_2^2$.

In particular, for any $\beta_1 < \beta_2$, $R(\beta_2) < R(\beta_1)$ and the high-probability stable-rank bound is tighter for the larger β .

Proof of Proposition 1. Let $h_k(s) = (1 - \beta_k)\beta_k^{t-s} \mathbf{1}_{s \leq t}$ denote the EMA impulse response. Two elementary spectral identities are used repeatedly (DC gain and noise gain, respectively):

$$\sum_{s \leq t} h_k(s) = 1 - \beta_k^t \xrightarrow{t \gg \tau_k} 1,$$

$$\sum_{s \leq t} h_k(s)^2 = (1 - \beta_k)^2 \frac{1 - \beta_k^{2t}}{1 - \beta_k^2} \xrightarrow{t \gg \tau_k} \frac{1 - \beta_k}{1 + \beta_k}.$$

The second collapse uses $\sum_{u \geq 0} \beta_k^{2u} = 1/(1 - \beta_k^2)$ together with $1 - \beta_k^2 = (1 - \beta_k)(1 + \beta_k)$.

Parallel component. Decompose $G^{s,\parallel} = \bar{G}^\parallel + \xi_\parallel^s$ with $\mathbb{E}[\xi_\parallel^s] = 0$. By linearity,

$$M_k^{t,\parallel} = \left(\sum_s h_k(s) \right) \bar{G}^\parallel + \sum_s h_k(s) \xi_\parallel^s,$$

so by (A2) (treating $\bar{G}^\parallel$ as time-stationary) the mean satisfies

$$\mathbb{E}[M_k^{t,\parallel}] = \left(\sum_s h_k(s) \right) \bar{G}^\parallel = (1 - \beta_k^t) \bar{G}^\parallel.$$

Since the two summands are orthogonal in L^2 (the second is zero-mean conditional on $\bar{G}^\parallel$), the second moment splits:

$$\mathbb{E}[\|M_k^{t,\parallel}\|_F^2] = (1 - \beta_k^t)^2 \|\bar{G}^\parallel\|_F^2 + \sum_{s,s'} h_k(s) h_k(s') \mathbb{E}[\langle \xi_\parallel^s, \xi_\parallel^{s'} \rangle].$$

For the fluctuation cross-term, (A3) gives $\mathbb{E}[\|\xi_\parallel^s\|_F^2] = \text{tr} \mathbb{E}[\xi_\parallel^s (\xi_\parallel^s)^\top] \leq r \sigma_\parallel^2$ since $\xi_\parallel^s \in \mathcal{S}$ has dimension r . Cauchy–Schwarz together with the lag- u autocorrelation bound $|\mathbb{E}[\langle \xi_\parallel^s, \xi_\parallel^{s'} \rangle]| \leq r \sigma_\parallel^2 |\rho(s - s')|$ then yields, by (A4),

$$\begin{aligned} & \sum_{s,s'} h_k(s) h_k(s') \mathbb{E}[\langle \xi_\parallel^s, \xi_\parallel^{s'} \rangle] \\ & \leq r \sigma_\parallel^2 \sum_{s,s'} h_k(s) h_k(s') |\rho(s - s')| \\ & \leq r \sigma_\parallel^2 C_\rho \sum_s h_k(s)^2 \\ & = \Theta\left(r C_\rho \sigma_\parallel^2 \frac{1 - \beta_k}{1 + \beta_k}\right), \end{aligned}$$

using $\sum_{s'} |\rho(s - s')| \leq 2C_\rho$ (Young’s inequality applied to $h_k \star |\rho|$). For $t \gg \tau_k$ the prefactor $(1 - \beta_k^t)^2 \rightarrow 1$ and we obtain

$$\mathbb{E}[\|M_k^{t,\parallel}\|_F^2] = \|\bar{G}^\parallel\|_F^2 + \Theta\left(r C_\rho \sigma_\parallel^2 \frac{1 - \beta_k}{1 + \beta_k}\right).$$

In the regime of interest the deterministic mean $\|\bar{G}^\parallel\|_F^2$ dominates the in-subspace fluctuation term (which is $O(1 - \beta_k)$ and vanishes as $\beta_k \rightarrow 1$), so

$$\mathbb{E}[\|M_k^{t,\parallel}\|_F^2] = \|\bar{G}^\parallel\|_F^2 (1 + o(1)).$$

Perpendicular component. The perpendicular projection has zero mean by (A1), so

$$\begin{aligned} \mathbb{E}[\|M_k^{t,\perp}\|_F^2] &= \sum_{s,s' \leq t} h_k(s) h_k(s') \\ &\quad \times \mathbb{E}[\langle G^{s,\perp}, G^{s',\perp} \rangle]. \end{aligned}$$

By Schwarz, (A3) and Cauchy–

$$\frac{\sqrt{\mathbb{E}[\|G^{s,\perp}\|_F^2] \cdot \mathbb{E}[\|G^{s',\perp}\|_F^2]}{\sigma_\perp^2} |\rho(|s-s'|)| \leq$$

$n\sigma_\perp^2 |\rho(|s-s'|)|$. Applying (A4),

$$\begin{aligned} \mathbb{E}[\|M_k^{t,\perp}\|_F^2] &\leq n\sigma_\perp^2 C_\rho \sum_s h_k(s)^2 \\ &= \Theta\left(n C_\rho \sigma_\perp^2 \frac{1-\beta_k}{1+\beta_k}\right). \end{aligned}$$

Energy ratio. Combining the two displays,

$$R(\beta_k) \triangleq \frac{\mathbb{E}[\|M_k^{t,\perp}\|_F^2]}{\mathbb{E}[\|M_k^{t,\parallel}\|_F^2]} = \kappa \frac{1-\beta_k}{1+\beta_k} (1+o(1)),$$

with $\kappa = n C_\rho \sigma_\perp^2 / \|\bar{G}^\parallel\|_F^2$. The map $\beta \mapsto (1-\beta)/(1+\beta)$ has derivative $-2/(1+\beta)^2 < 0$, so $R(\beta_k)$ is strictly decreasing on $[0, 1)$ and tends to 0 as $\beta_k \rightarrow 1^-$.

High-probability stable-rank envelope. Let $X^s = h_k(s) G^{s,\perp}$, so $M_k^{t,\perp} = \sum_s X^s$. Under (A1)–(A4), the two-sided spectral-norm bounds in (A3) (row- and column-variance) and the boundedness $\|G^{s,\perp}\|_2 \leq L$ a.s. (which holds with gradient clipping at norm L), the matrix Bernstein inequality (Tropp, 2015) gives, for any $\delta \in (0, 1)$,

$$\begin{aligned} \Pr\left[\|M_k^{t,\perp}\|_2 \geq C_1 \sigma_\perp \sqrt{\frac{1-\beta_k}{1+\beta_k} \log \frac{2(m+n)}{\delta}} \right. \\ \left. + C_2 L (1-\beta_k) \log \frac{2(m+n)}{\delta}\right] \leq \delta, \end{aligned}$$

with absolute constants C_1, C_2 depending only on C_ρ . Thus the perpendicular spectral norm shrinks like $\sqrt{(1-\beta_k)/(1+\beta_k)}$ up to logs, while the parallel spectral norm remains $\Theta(\|\bar{G}^\parallel\|_2)$. Consequently

$$\text{srnk}(M_k^t) \leq \text{srnk}(\bar{G}^\parallel) + o(1) \leq r$$

as $\beta_k \rightarrow 1^-$, with probability at least $1 - \delta$. $\square$

Empirical methodology. Figure 2 reports per-index singular values of the two momentum buffers from the final-step checkpoint ($t=10,172$) of the $K=2$ Multiscale Muon run, computed over the 12 square 1024×1024 weight matrices in that checkpoint and shown as median with 25–75% interquartile range; the buffer convention matches Section 5.1 ($\beta_1=0.969$, $\tau_1=32$, wide-band; $\beta_2=0.998$, $\tau_2=512$, narrow-band).

Relationship to Proposition 1’s assumptions. The singular-value ordering in Figure 2 validates the *conclusion* of Proposition 1 and provides indirect evidence for its two structural assumptions (slowly-rotating low-rank parallel subspace;

weakly autocorrelated zero-mean perpendicular component). The slow-rotation assumption is supported by Zhao et al. (2024), who report that the dominant GaLore subspace evolves slowly on LLaMA-scale pretraining (so a periodic basis refresh suffices); we use a similar rank ratio and refresh schedule ($r/n=0.125$, $T=200$). The weak-autocorrelation assumption is standard for stochastic-gradient noise. Were either assumption to fail badly, the predicted ordering $\text{srnk}(M_2) \leq \text{srnk}(M_1)$ would not hold across layers as it does in Figure 2.

Direct verification on the full-rank $K=3$ run.

The main text (Figure 2) verifies Proposition 1’s two assumptions on the operational S-MuG $K=2$ checkpoint stream. Figure 4 repeats the analysis on the full-rank Multiscale Muon $K=3$ run and shows the same β -monotone pattern: at $T=100$, overlap increases with β ($0.516 \rightarrow 0.526 \rightarrow 0.699$ for $\beta=0.969, 0.992, 0.998$), and the momentum cosine does likewise ($-0.03 \rightarrow 0.14 \rightarrow 0.58$). At long lag ($T=5000$), overlap remains above baseline for all three buffers ($0.481, 0.456, 0.393$ vs. 0.125), while cosine decays toward zero with a clear ordering ($0.003, 0.013, 0.038$), confirming the slow-subspace structure exploited by S-MuG is not an artifact of rank truncation.

D 1 B and 3 B Downstream Evaluation

Table 10 reports the per-task zero-/few-shot scores referenced in Section 5.5 for the 1 B model (mean $\pm$ std over $n=2$ seeds $\{42, 43\}$; values in %).

Discussion. Three patterns hold robustly across both seeds. (i) *Code metrics clearly favour the multiscale variants over single Muon, and S-MuG matches full-rank Multiscale Muon at $1.25\times$ the optimizer memory:* Code average is $7.9 \pm 1.1\%$ for Muon vs. $9.3 \pm 0.7\%$ (MS Muon) and $9.2 \pm 0.8\%$ (S-MuG); the single-Muon $\rightarrow$ multiscale gain on code is much larger than the multiscale $\rightarrow$ S-MuG step, suggesting most of the code improvement comes from the second buffer itself rather than from how it is compressed. (ii) *Reasoning is essentially flat across all three optimizers* ($45.1/45.1/45.5\%$, all within ± 0.6); per-task variation is within 1–2 points and the leader differs by task (Muon wins WG/BoolQ, S-MuG wins OBQA/ARC-e/HS/PIQA, MS Muon wins ARC-c), which is what one expects when per-task seed

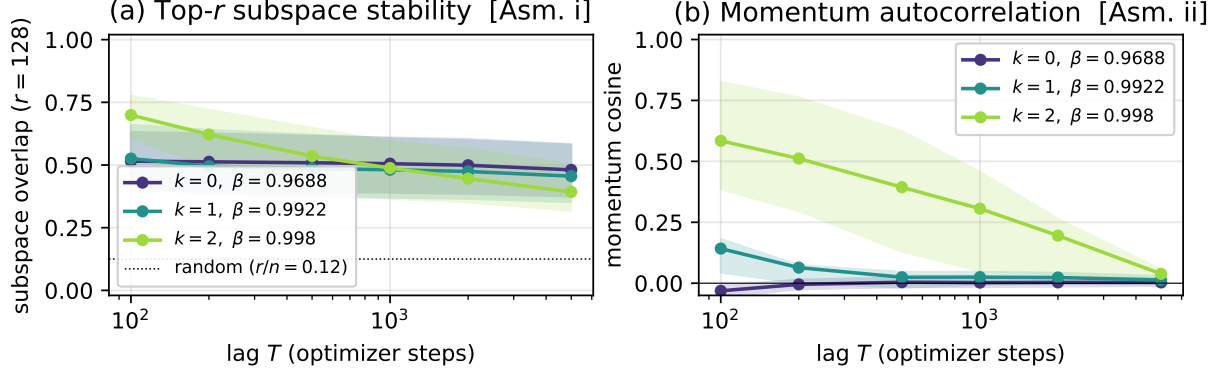


Figure 4: Direct verification of Proposition 1’s assumptions on the full-rank Multiscale Muon $K=3$ run (same protocol as Figure 2: 86 checkpoints over $[500, 9000]$, IQR over all $(t, t+T)$ pairs and 1024×1024 parameters). (a) top- $r=128$ subspace overlap; (b) momentum cosine. The β -monotone separation matches the $K=2$ S-MuG result.

Table 10: 1 B GQA, downstream zero-/few-shot evaluation at iter 10 000 (mean $\pm$ std over $n=2$ seeds $\{42, 43\}$; values in %). HE = HumanEval, MB = MBPP (both reported as pass@1/pass@10); WG = WinoGr., HS = HellaSw. Code avg. = mean of the two pass@1 code scores (HE p@1, MB p@1); Reason avg. = mean of the seven QA/commonsense tasks (OBQA, ARC-e, ARC-c, WG, HS, PIQA, BoolQ). Overall benchmark average (cited in the main text) likewise uses HE/MB pass@1 only.

Optimizer	HE	MB	OBQA	ARC-e	ARC-c	WG	HS	PIQA	BoolQ	Code	Reason
Muon ($K=1$)	3.0 $\pm$ 0.2/8.3 $\pm$ 2.5	12.8 $\pm$ 2.3/23.9 $\pm$ 3.6	28.6 $\pm$ 0.8	46.8 $\pm$ 1.4	24.4 $\pm$ 0.1	52.4 $\pm$ 0.6	37.0 $\pm$ 0.1	65.6 $\pm$ 0.7	61.2 $\pm$ 0.8	7.9 $\pm$ 1.1	45.1 $\pm$ 0.3
MS Muon ($K=2$)	4.2 $\pm$ 0.7/9.4 $\pm$ 1.2	14.3 $\pm$ 0.7/24.8 $\pm$ 0.9	28.8 $\pm$ 1.7	47.4 $\pm$ 0.5	26.3 $\pm$ 0.7	51.7 $\pm$ 0.1	36.8 $\pm$ 0.3	65.9 $\pm$ 0.4	58.4 $\pm$ 4.2	9.3 $\pm$ 0.7	45.1 $\pm$ 0.3
S-MuG ($K=2$)	4.2 $\pm$ 0.8/7.6 $\pm$ 0.2	14.1 $\pm$ 0.8/25.8 $\pm$ 0.3	29.8 $\pm$ 2.0	47.9 $\pm$ 0.1	24.6 $\pm$ 0.1	52.5 $\pm$ 1.1	37.3 $\pm$ 0.2	66.0 $\pm$ 0.3	60.7 $\pm$ 1.5	9.2 $\pm$ 0.8	45.5 $\pm$ 0.6

scatter at the 1 B scale is itself $\sim 1\text{--}3$ pt. (iii) *The downstream picture is consistent with the LM-PPL story* (Table 5): the ordering $\text{S-MUG} \geq \text{MS Muon} \geq \text{Muon}$ on overall benchmark average ($37.5 \pm 0.7/37.1 \pm 0.4/36.9 \pm 0.4\%$) mirrors the same ordering on every LM-loss column and holds in both seeds individually. The takeaway from Appendix D is therefore not “S-MUG is uniformly best on every per-task metric” but rather “selectively compressing the narrow-band buffer preserves the multiscale gain on both LM loss and aggregate downstream while costing $4\times$ less extra optimizer memory than full-rank MS Muon”.

3 B downstream details. Table 11 gives the detailed Code and Reasoning results for the 3 B scale-transfer evaluation referenced in Section 5.5. Values are from iteration 30,000 and use one seed per optimizer; Code and Reason are averages defined as in Table 10.

The detailed metrics show consistent advantages for S-MUG: it out-performs single-scale Muon on both aggregate Code (18.06% vs. 17.03%) and Reasoning (50.29% vs. 49.15%) averages, as well as on a majority of individual tasks (including OBQA, ARC-easy, ARC-challenge, HS, PIQA,

and BoolQ). This confirms that selective multiscale momentum successfully preserves and translates its optimization advantages to downstream capacities at the 3 B parameter scale.

E Wall-clock Comparison

End-to-end wall-time ($8 \times \text{MI300X}$, 10,172 steps).

All rows in Table 3 use identical timing instrumentation: the clock is paused around validation, checkpoint I/O, and any in-loop diagnostics, so the reported hours measure pure forward+backward+optimizer cost. Measured per-step averages give Muon $K=1$ 1.154 s, MS Muon $K=2$ 1.160 s, MS Muon $K=3$ 1.163 s, S-MUG $K=2$ 1.176 s, S-MUG $K=3$ 1.164 s, AdamW 1.177 s. Two takeaways: (i) S-MUG adds a per-step rank- r GaLore projection on the narrow-band buffer plus a periodic truncated-SVD refresh, costing +1.38% over full-rank Multiscale Muon at $K=2$ and +0.09% (within noise) at $K=3$: at $K=3$ the projection cost is fully absorbed by the saved Newton-Schulz on the compressed branch. (ii) AdamW is *not* cheaper than Muon at this scale (forward/backward dominates and the AdamW state update is memory-bandwidth bound); it is essen-

Table 11: 3 B downstream Code and Reasoning evaluation at iteration 30,000 (one seed per optimizer; values in %). HE = HumanEval, MB = MBPP, WG = WinoGr., HS = HellaSw. Code is the mean of HE/MB pass@1; Reason is the mean of the seven reasoning tasks.

Optimizer	HE	MB	OBQA	ARC-e	ARC-c	WG	HS	PIQA	BoolQ	Code	Reason
Muon	11.43/22.45	22.63/37.38	31.00	55.56	28.58	55.09	47.11	68.55	58.13	17.03	49.15
S-MUG	10.95/21.21	25.17/38.76	33.40	56.94	30.29	53.43	48.02	69.04	60.89	18.06	50.29

tially tied with S-MUG $K=2$.

F Additional Ablations

F.1 NS Placement and Rank

S-MUG ($K=2$) differs from full-rank Multiscale Muon along two axes simultaneously: (i) the wide-band buffer is rank- $r=128$ truncated rather than kept full-rank, and (ii) one Newton–Schulz (NS) pass is applied to the combined band update $s_1 M_1 + s_2 M_2 V^\top$ rather than per-buffer-then-summed (as noted in Limitations and in Pappone and Selvaraj, 2025). To attribute the observed quality preservation specifically to the selection rule rather than to the NS placement, we report the missing two off-diagonal cells of the 2×2 design at 201 M / 5 B tokens (single seed; the diagonal cells are the multi-seed Multiscale Muon $K=2$ and S-MUG $K=2$ entries of Table 3).

Table 12: NS placement $\times$ rank ablation at 201 M (PPL $\downarrow$). Diagonal cells are the multi-seed Table 3 entries (MS Muon $K=2$ top-left; S-MUG $K=2$ bottom-right); off-diagonal cells are single-seed NEW runs. Combined- and per-buffer NS are statistically tied when both buffers are full-rank, but combined-NS *wins* once the narrow-band buffer is rank-truncated.

rank on β_2 (narrow)	per-buf NS (sum-of-NS)	combined NS (NS-of-sum)
full ($r_2=n=1024$)	21.69 $\pm$ 0.03	21.71
$r_2=128$ (GaLore)	21.73	21.66$\pm$0.05

Both off-diagonal cells sit within 0.07 PPL of the diagonal and inside the seed scatter ($\sigma \sim 0.05$ PPL for S-MUG), so the selection rule itself accounts for the headline quality preservation. But the off-diagonal numbers also expose an *interaction* that justifies the combined-NS design: with both buffers full-rank, per-buf and combined are statistically tied (21.69 vs. 21.71); once the narrow-band buffer is rank-128 truncated, combined-NS overtakes per-buf by 0.07 PPL (21.66 vs. 21.73). This is the qualitative pattern predicted by the argument that orthogonalising a rank- r low-rank surrogate in isolation (per-buf NS on $M_2 V^\top$ when

M_2 is held in $r \times n$ form) discards the corrective full-space content carries by M_1 , whereas summing first produces a numerically full-rank update before NS sees it.

F.2 Multi-Band K -Sweep

Table 13 reports the broader K -sweep for full-rank Multiscale Muon at 201 M. Adding a second band gives the main improvement, a third band is essentially neutral, and $K=4/5$ slightly regress. Since $\beta = 1 - 1/\tau$, the $K=5$ wide-range setting spans $\tau = 32\text{--}2048$ with widely separated time constants, whereas the narrow-range setting packs five more densely spaced values into $\tau = 32\text{--}512$; neither improves over the $K=2/3$ design point in this budget. The $K \leq 3$ entries use $n=3$ seeds, while $K \geq 4$ uses one seed.

Table 13: Multi-band K -sweep for full-rank Multiscale Muon at 201 M. $n=3$ seeds for $K \leq 3$; $n=1$ for $K \geq 4$.

K	Val. PPL $\downarrow$
1	21.81 $\pm$ 0.02
2	21.69 $\pm$ 0.03
3	21.68 $\pm$ 0.02
4	21.73
5 (wide)	21.76
5 (narrow)	21.72

G Spectral Score Diagnostic and Empirical λ -sweep Curves

Spectral score $g(\omega)$ fit. We fit a piecewise-constant non-negative score $g(\omega)$ ($N_{\text{bands}}=5$, log-frequency measure) jointly against the $K=1$ single-buffer sweep and the $K=2$ (τ_1, τ_2, λ) $\rightarrow$ val loss table. Figure 5(a) shows the recovered $g(\omega)$ (positive low-frequency lobe, sign change at mid-frequency); panel (b) confirms it ranks $K=1$ runs consistently with observed val loss; panel (c) shows that the resulting λ^* predictions correlate with observed optima across (τ_1, τ_2) pairs (Spearman $\rho=0.97$, MAE 0.12). We stress that this fit is post-hoc on the same (τ_1, τ_2) family that it is then evaluated on, and Appendix H shows that no non-negative linear spectral

score on equal-energy normalised filters can produce an interior λ^* in closed form; the diagnostic is therefore a consistency check on the filter-bank picture, not an independent predictor.

Empirical λ -sweep curves. Figure 6 reports the final validation PPL (10 172 steps, single seed) as a function of the narrow-band mixing weight λ for five (τ_1, τ_2) pairs that were sampled at ≥ 5 values of λ in our $K=2$ multiscale Muon sweep ($\tau_1 = \text{wide-band}$, $\tau_2 = \text{narrow-band}$). Each curve is U-shaped with an interior optimum, in agreement with Section 5.2; the optimum λ^* shifts with (τ_1, τ_2) (e.g. $\lambda^* \approx 0.5$ at (16, 64) but $\lambda^* \approx 0.1$ at (32, 256)). The interior optima and their bandwidth dependence are precisely what Appendix H rules out from any non-negative linear spectral score. Two practical takeaways: (i) at any fixed (τ_1, τ_2) the PPL surface is shallow, with typical curvature near the minimum of ≤ 0.2 PPL per $\Delta\lambda=0.2$, so a 3–5 point grid is sufficient at a new scale; (ii) widely-separated (τ_1, τ_2) such as (16, 256) or (64, 256) favor very small λ , while moderately-separated pairs such as (16, 64) favor $\lambda \approx 1/2$, consistent with the filter-bank picture (Section 3) that the wide-band branch dominates when the two bandwidths are dissimilar.

H Why a Linear Spectral Score Cannot Predict λ^*

This appendix records a short observation that motivates our *empirical* (rather than closed-form) fit of $g(\omega)$ in Section 5.2 and explains why the interior λ^* documented empirically in Appendix G cannot be predicted by any non-negative linear spectral score: under the equal-energy normalisation of the EMA filters used throughout the paper, no linear mixture of two EMA filters can produce a non-trivial interior optimum in λ via any non-negative spectral weight.

Setup. Let $\hat{H}_\beta(\omega) = \sqrt{1 - \beta^2}/(1 - \beta e^{-j\omega})$ be the equal-energy EMA filter, and write the mixed filter as

$$\hat{H}_{\text{combo}}(\omega, \lambda) = \lambda \hat{H}_{\beta_s}(\omega) + (1 - \lambda) \hat{H}_{\beta_f}(\omega), \quad (5)$$

with $\lambda \in [0, 1]$. For any non-negative spectral weight $W(\omega) \geq 0$ and positive measure $d\mu$ on $[-\pi, \pi]$, define the score

$$P_W(\lambda) = \int W(\omega) |\hat{H}_{\text{combo}}(\omega, \lambda)|^2 d\mu. \quad (6)$$

Expanding the modulus gives the quadratic form

$$P_W(\lambda) = A_W \lambda^2 + B_W (1 - \lambda)^2 + 2C_W \lambda(1 - \lambda), \quad (7)$$

with coefficients

$$\begin{aligned} A_W &= \int W |\hat{H}_{\beta_s}|^2 d\mu, \\ B_W &= \int W |\hat{H}_{\beta_f}|^2 d\mu, \\ C_W &= \int W \text{Re}(\hat{H}_{\beta_s} \hat{H}_{\beta_f}^*) d\mu. \end{aligned} \quad (8)$$

Lemma 1 (Cross-term identity). $C_W = \rho(A_W + B_W)$ with $\rho = \sqrt{(1 - \beta_s^2)(1 - \beta_f^2)}/[2(1 - \beta_s\beta_f)] \in (0, \frac{1}{2}]$, independent of W and $d\mu$.

Proposition 2 (Linear-score convexity rules out an interior maximiser). For any $W \geq 0$ and any positive measure $d\mu$, $P_W(\lambda)$ is convex on $[0, 1]$. Its maximum on $[0, 1]$ is therefore always attained at an endpoint, so no non-negative linear spectral score (i.e. a single positive-weight integral of $|\hat{H}_{\text{combo}}|^2$) can place its maximiser in the open interval $\lambda \in (0, 1)$.

Proof of Lemma 1. Let $D(\omega) = (1 - 2\beta_s \cos \omega + \beta_s^2)(1 - 2\beta_f \cos \omega + \beta_f^2) > 0$ be the common denominator of $|\hat{H}_{\beta_s}(\omega)|^2$, $|\hat{H}_{\beta_f}(\omega)|^2$, and $\text{Re}(\hat{H}_{\beta_s} \hat{H}_{\beta_f}^*)$. Direct computation gives the pointwise identities

$$\begin{aligned} |\hat{H}_{\beta_s}|^2 + |\hat{H}_{\beta_f}|^2 &= \frac{2(1 - \beta_s\beta_f) N(\omega)}{D(\omega)}, \\ 2 \text{Re}(\hat{H}_{\beta_s} \hat{H}_{\beta_f}^*) &= \frac{2\sqrt{(1 - \beta_s^2)(1 - \beta_f^2)} N(\omega)}{D(\omega)}, \end{aligned}$$

with $N(\omega) = (1 + \beta_s\beta_f) - (\beta_s + \beta_f) \cos \omega \geq 0$ (since $\beta_s, \beta_f \in [0, 1]$). Their pointwise ratio is therefore the constant 2ρ with $\rho = \sqrt{(1 - \beta_s^2)(1 - \beta_f^2)}/[2(1 - \beta_s\beta_f)]$, independent of ω , W , and $d\mu$; integrating against $W(\omega) d\mu$ yields $C_W = \rho(A_W + B_W)$. For the bound $\rho \leq \frac{1}{2}$, AM–GM gives $1 - \beta_s\beta_f \geq \sqrt{(1 - \beta_s^2)(1 - \beta_f^2)}$ (equivalent to $(\beta_s - \beta_f)^2 \geq 0$ after squaring), so $\rho \leq \frac{1}{2}$ with equality iff $\beta_s = \beta_f$. $\square$

Proof of Proposition 2. Substituting $C_W = \rho(A_W + B_W)$ into the quadratic and collecting terms,

$$\begin{aligned} P_W(\lambda) &= (A_W + B_W)(1 - 2\rho)\lambda^2 \\ &\quad + 2[\rho(A_W + B_W) - B_W]\lambda + B_W. \end{aligned}$$

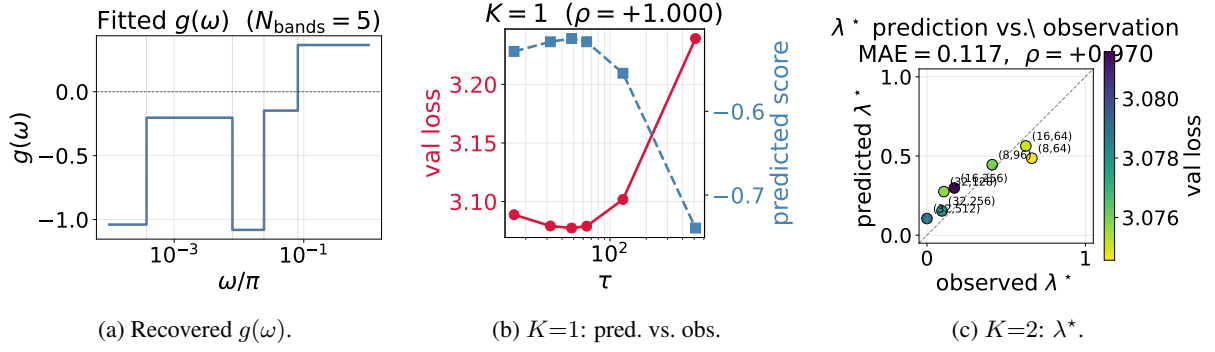


Figure 5: Spectral score $g(\omega)$ fit ($N_{\text{bands}}=5$, log measure) is consistent with the empirical optima. (c) Predicted λ^* correlates with observed optimum across (τ_1, τ_2) pairs (Spearman $\rho=0.97$, MAE 0.12). Diagnostic only; not a closed-form recipe (Appendix H).

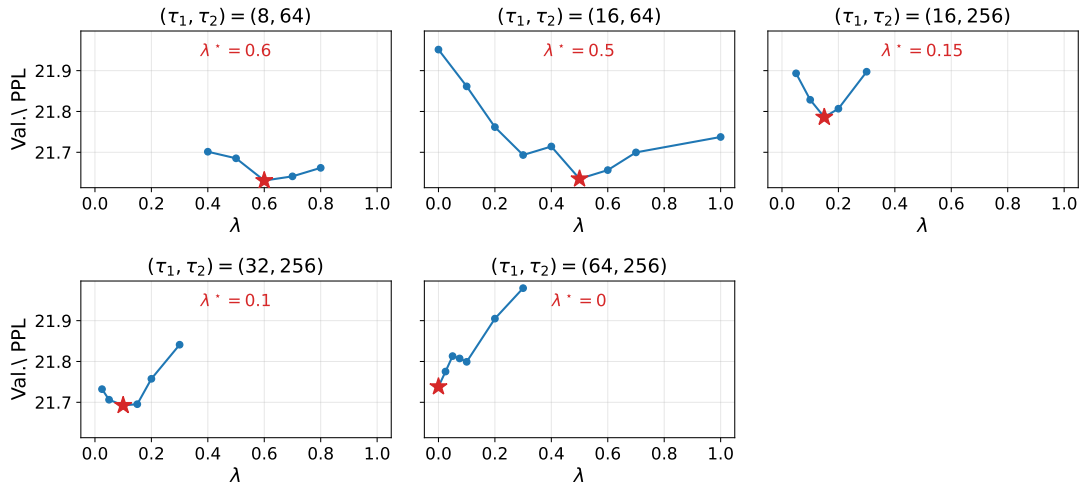


Figure 6: Final validation PPL vs. narrow-band mixing weight λ for five $K=2$ (τ_1, τ_2) combinations from our multiscale Muon sweep (τ_1 = wide-band, τ_2 = narrow-band; identical 201 M setup as Section 5). Red stars mark each curve’s empirical minimum. The interior optimum and its shift with (τ_1, τ_2) are the empirical content of the bandwidth dependence (Section 3); the shallowness of each curve means a coarse 3–5-point grid is enough to find λ^* at a new scale.

The coefficient of λ^2 is $(A_W + B_W)(1 - 2\rho) \geq 0$ by Lemma 1 (with strict inequality unless $\beta_s = \beta_f$), so P_W is convex on $[0, 1]$ and any maximum on $[0, 1]$ is attained at an endpoint $\lambda \in \{0, 1\}$. Equivalently, the unique interior critical point $\lambda^* = (B_W - \rho(A_W + B_W)) / [(A_W + B_W)(1 - 2\rho)]$, when it lies in $(0, 1)$, is a *minimiser* of P_W , never a maximiser. $\square$

What this means. A higher spectral score $P_W(\lambda)$ models better filter quality; thus a PPL interior *minimiser* at $\lambda^* \in (0, 1)$ corresponds to an interior *maximiser* of P_W . Proposition 2 shows that no non-negative W can produce an interior maximiser of P_W (the maximum is always at an endpoint), so no non-negative linear spectral score can predict the interior PPL minimum. Equiva-

lently, the interior λ^* minimum of validation PPL observed in Appendix G would require a surrogate score that is *concave* (or at least non-convex) in λ near the optimum. Proposition 2 shows the entire non-negative linear-spectral class is convex in λ , so an interior λ^* must arise from one of: (i) the per-buffer Newton–Schulz nonlinearity, which breaks the quadratic structure of $|\hat{H}_{\text{combo}}|^2$; (ii) genuinely non-quadratic geometry of the loss; or (iii) a spectral weight that is *allowed to change sign*. Our data-dependent fit of $g(\omega)$ in Section 5.2 takes route (iii): the recovered g exhibits a sign change near mid-frequency (Figure 5a), which is precisely the structure Proposition 2 rules out for non-negative weights.