MuLoRA: Spectrally Balanced Low-Rank Adaptation for Continual Learning

Anonymous authors
Paper under double-blind review

Abstract

Low-rank adaptation (LoRA) provides a parameter-efficient approach to continual learning, but its nominal rank can conceal a loss of effective adaptation capacity. We identify *spectral plasticity collapse*: during sequential adaptation, update energy becomes concentrated in a small subset of singular modes, leaving much of the available low-rank space underutilized. This exposes a limitation of interference avoidance alone: protecting historical representations does not ensure that the remaining adaptation capacity is responsive to new tasks or effectively utilized. To address this problem, we propose `MuLoRA`, which jointly controls capacity allocation and utilization. First, historical whitening identifies input directions with strong current-task response relative to accumulated historical response, yielding a task-adaptive basis that remains fixed during training. Second, approximate polar orthogonalization of momentum updates reduces spectral concentration within the selected space. An orthonormal basis connects these mechanisms by transferring the factor-update spectrum exactly to the induced weight update. We establish a max–min characterization of exact subspace selection and derive cumulative spectral bounds under controlled cross-step anisotropy. Across five class-incremental benchmarks and eight incremental settings, `MuLoRA` achieves the highest mean accuracy in 15 of 16 reported metrics.

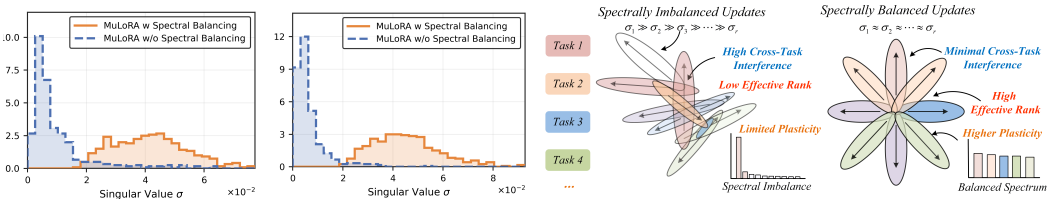
1 Introduction

Continual learning (CL) requires models to acquire knowledge from sequential tasks while retaining performance on previous ones Kirkpatrick et al. (2017); Qiu et al. (2025b); Wang et al. (2021). With the emergence of foundation models, this challenge increasingly concerns adapting pretrained representations under limited computation and memory budgets Radford et al. (2021); Cheng et al. (2024). Low-Rank Adaptation (LoRA) Hu et al. (2022) provides an efficient parameterization by expressing each task-specific weight update as $\Delta W = BA$ while freezing the backbone within each task. However, a fixed adaptation rank does not ensure that all available directions are effectively used. An update can retain its algebraic rank while concentrating most of its energy in only a few singular modes. This distinction matters in continual adaptation, where a limited low-rank space must support learning across successive tasks.

We observe a phenomenon that exposes this limitation: ***spectral plasticity collapse***. During continual LoRA adaptation, the learned update ΔW can exhibit a highly concentrated spectrum, leaving many available modes weakly utilized (Figure 1 and Figure 6). We examine this behavior through four complementary diagnostics: the singular-value distribution, effective rank, the nonzero condition number κ_+, and stable rank. The former reveals how update energy is distributed across modes, while the latter quantifies the effective number of modes carrying that energy. Here, collapse refers to the concentration of update energy in a few singular directions, leading to lower effective and stable ranks without necessarily reducing the algebraic rank. Thus, maintaining the nominal LoRA rank does not guarantee effective use of its adaptation capacity.

Muon Jordan et al. (2024) offers a natural starting point for addressing this spectral concentration. Through approximate polar orthogonalization of momentum, it reshapes the update spectrum toward balanced nonzero singular values Bernstein (2025). This motivates using spectral balancing to improve update utilization within a fixed-rank adapter. However, spectral balancing of LoRA factors does not automatically translate into spectral balancing of the weight update. When both factors are trained,

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.



(a) Task 1 **(b)** Task 20 **(c)** Spectral Capacity Utilization

Figure 1: **Spectral concentration limits effective low-rank capacity.** (a,b) Singular-value distributions (Frequency vs. σ) of learned LoRA updates ΔW at Tasks 1 and 20, with and without spectral balancing. More balanced spectra distribute update energy across a larger number of modes. (c) Conceptual illustration of concentrated versus balanced spectra and their relationship to effective rank. A fixed nominal rank can correspond to markedly different levels of spectral utilization.

their changes induce $\delta W = \delta B A + B \delta A + \delta B \delta A$. Even if the factor steps are individually balanced, multiplication by the factors and interaction between terms can produce an uneven weight-step spectrum. The challenge is to make Muon's spectral control carry to update of the model.

We address this issue by fixing a row-orthonormal basis A and updating only B with Muon. Since $AA^\top = I_r$, each weight step $\delta W = \delta B A$ has the same nonzero singular values as δB. Muon's spectral balancing therefore transfers directly to the weight step. Likewise, BA preserves the nonzero singular values of B, so the basis introduces no spectral distortion.

Fixing A makes the choice of its subspace important, because the model can only update within that space. An orthonormal basis preserves the update spectrum, but its directions may not be useful for the new task. Methods such as InfLoRA Liang & Li (2024) avoid a retained historical feature subspace to reduce interference in Figure 2. However, shared directions can also support new learning, and directions with little historical response may still be unhelpful for the current task. We therefore select directions with strong current-task response relative to their historical response, providing a task-aware space for spectral balancing.

These observations suggest two connected requirements. The adaptation space should contain directions with strong current-task response relative to historical exposure, and its parameterization should preserve spectral control from factor space to weight space. Together, they lead to following question:

How can Muon's spectral balancing be transferred to LoRA weight updates within a subspace suited to continual adaptation?

We propose Muon Low-Rank Adaptation (MuLoRA), which connects Relative-Response Spectral Allocation (RRS) with Subspace Spectral Balancing (SSB). For each task, MuLoRA constructs a row-orthonormal input basis A_t, keeps it fixed, and applies Muon-style updates to the trainable up-projection B_t. The central design principle is to select task-responsive directions using historical information and preserve the update spectrum through the LoRA parameterization.

For subspace selection, we maximize current-task feature response under a regularized historical-response budget. Historical whitening converts this objective into a standard spectral selection problem. The selected directions are mapped back to the input space and orthonormalized to construct A_t. Historical information thus provides a graded selection metric: previously active directions remain eligible when their relative current-task response is strong. With exact spectral selection, the resulting subspace maximizes the worst-direction relative response among all subspaces of the same dimension. This provides a subspace-wide characterization of the directions available for adaptation.

For spectral balancing, we apply approximate polar orthogonalization to momentum updates of B_t. The ideal transformation equalizes all nonzero singular values, maximizing spectral participation on the active update support. Since $A_t A_t^\top = I_r$, the resulting factor step and its induced weight step have identical nonzero singular values and effective rank. This connects the two components: RRS determines where adaptation occurs and supplies an orthonormal basis, while SSB controls how energy is distributed in each weight step.

Our analysis distinguishes this stepwise control from the behavior of accumulated updates. Balanced steps need not produce a balanced sum, because successive steps can reinforce or cancel different

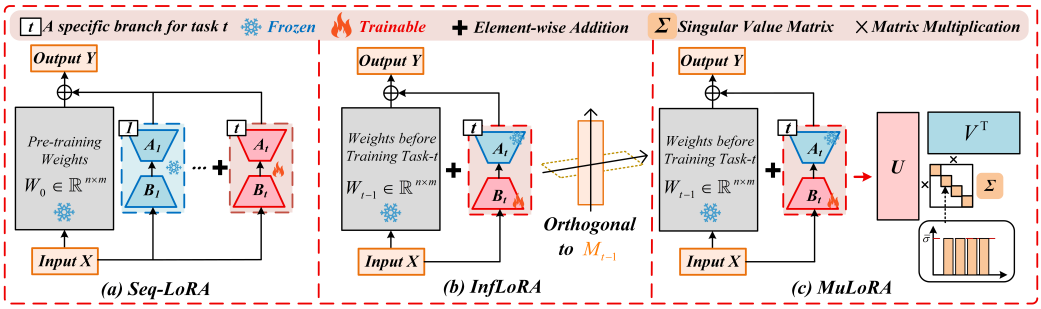


Figure 2: **Comparison of different LoRA-based CL paradigms.** (a) Seq-LoRA employs independent adapters, (b) InfLoRA preserve previous knowledge by restricting updates to isolated directions, (c) MuLoRA fixes a task-responsive orthonormal basis and applies Muon to the trainable factor, transferring spectral balancing directly to weight steps.

directions. We derive cumulative condition-number and spectral participation bounds under controlled cross-step anisotropy, identifying conditions under which stepwise balancing extends to accumulated adaptation. Empirically, we examine singular-value distributions and participation-based effective rank alongside continual-learning accuracy. Across five benchmarks and eight incremental settings, MuLoRA achieves the highest mean accuracy in 15 of 16 metrics. Ablations support jointly selecting task-responsive subspaces and balancing updates within them.

Our main contributions can be summarized as follows:

- We identify *spectral plasticity collapse* in continual LoRA adaptation: update energy concentrates in a few singular modes, leaving the effective rank well below the nominal adapter rank. We further highlight a challenge in applying Muon: balancing the two factor steps separately does not guarantee a spectral balanced weight step.
- We propose MuLoRA, a history-aware low-rank parameterization that transfers Muon's spectral balancing directly to weight space. It fixes an orthonormal input basis selected by current-task response relative to historical exposure and applies Muon updates to the trainable factor. This design preserves the nonzero singular values of each factor step in the induced weight step, enabling spectral control within a task-responsive adaptation space.
- We show that exact subspace selection maximizes the worst-direction relative response and derive cumulative spectral bounds under controlled cross-step anisotropy. Across five benchmarks and eight incremental settings, MuLoRA achieves the highest mean accuracy in 15 of 16 reported metrics, with ablations supporting the complementary benefits of RRS and SSB.

2 Related Works

PEFT-based Continual Learning. Parameter-efficient fine-tuning (PEFT) enables continual adaptation of pretrained models through lightweight trainable modules. Prompt-based methods Smith et al. (2023); Gao et al. (2024) learn and select prompts to accommodate sequential tasks, while PGP Qiao et al. (2023) and VPT-NSP Lu et al. (2024) derive projection constraints for mitigating interference. Adapter-based methods Li et al. (2025); Tan et al. (2024); He et al. (2025); Wang et al. (2026) preserve and integrate task knowledge through lightweight modules, representation alignment, or feature compensation. LoRA-based approaches further structure adaptation within low-rank parameter spaces. SD-LoRA Wu et al. (2025b) decouples direction and magnitude learning; O-LoRA Wang et al. (2023) and InfLoRA Liang & Li (2024) restrict overlap with historical subspaces to reduce interference. SplitLoRA Qiu et al. (2025a), BiLoRA Zhu et al. (2025), and PLAN Wang et al. (2025) further explore subspace decomposition, orthogonality, and allocation to balance stability and plasticity. Complementing these approaches, EWC-LoRA Zheng et al. (2026) regularizes a shared low-rank update using parameter importance estimated in the full weight space. Closely related to our work, EBLoRA Gu et al. (2026) links spectral imbalance to forgetting and enforces equal nonzero singular values through orthonormal factors and a shared scalar magnitude. MuLoRA addresses the stability–plasticity trade-off by selecting subspaces with strong current-task response relative to historical exposure and balancing momentum steps within them.

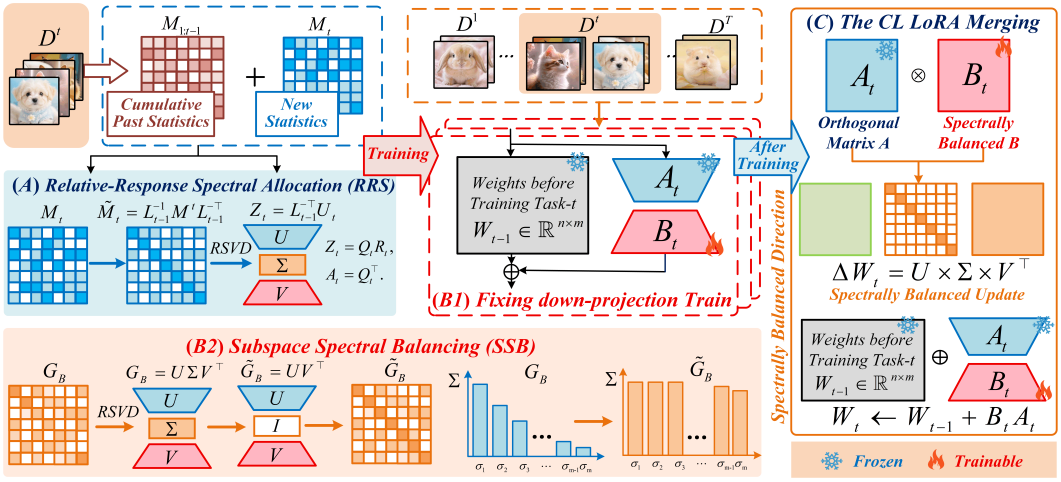


Figure 3: **Relative-response subspace selection and spectrally controlled optimization.** (A) Historical whitening identifies directions with strong current-task response relative to accumulated historical response, from which we construct a row-orthonormal basis A_t. (B1) We freeze the backbone and A_t, and train only B_t. (B2) Approximate polar orthogonalization reshapes the momentum spectrum, balancing individual optimization steps within the selected subspace. (C) After each task, the learned update is merged into the backbone as $W_t = W_{t-1} + B_t A_t$.

3 Methodology

Problem Definition. In CL, there are T sequential tasks with non-overlapping classes. The datasets are defined as $\{\mathcal{D}^1, \cdots, \mathcal{D}^t, \cdots, \mathcal{D}^T\}$. $\mathcal{D}^t = \{x_i^t, y_i^t\}_{i=1}^{N^t}$ denotes the t-th task's dataset, where $x_i^t \in \mathbb{R}^{H\times W\times C}$ is the i-th input and $y_i^t \in \mathcal{C}^t$ is its class label. $\mathcal{C}^t$ denotes the label space of the t-th task, and $\mathcal{C}^t \cap \mathcal{C}^{t'} = \emptyset$ for $t \neq t'$. The objective of CL is to train a model $f(w, \cdot)$ sequentially on T tasks and performs well on all seen classes $\mathcal{C}^1 \cup \cdots \cup \mathcal{C}^T$. We consider the task-agnostic class-incremental learning setting, where task identity is unavailable during inference.

Overall framework. Figure 3 illustrates the proposed framework. Following existing PEFT-based methods Wang et al. (2022b); Tan et al. (2024), we adopt a pretrained Vision Transformer (ViT) Dosovitskiy et al. (2021) as the backbone. For task t, each adapted weight matrix $W_{t-1} \in \mathbb{R}^{n\times m}$ remains frozen and is augmented with a rank-r update $B_t A_t$, where $A_t \in \mathbb{R}^{r\times m}$ and $B_t \in \mathbb{R}^{n\times r}$. Historical whitening identifies a task-adaptive subspace with strong current-task response relative to accumulated historical response. We construct a row-orthonormal basis A_t of this subspace, satisfying $A_t A_t^\top = I_r$, and keep it fixed throughout the task. Starting from $B_t = 0$, we optimize B_t using Muon-style approximate polar orthogonalization of momentum updates. The orthonormal basis preserves the nonzero singular values between each factor update and its induced weight update, connecting subspace selection with stepwise spectral control. For an input feature $x \in \mathbb{R}^m$, the adapted layer computes $y = (W_{t-1} + B_t A_t)x \in \mathbb{R}^n$. After training, we merge the learned update into the backbone as $W_t = W_{t-1} + B_t A_t$ and construct a new basis for the next task.

3.1 Relative-Response Spectral Allocation

Motivation: shared feature directions also can support new learning. Isolation-based methods such as InfLoRA restrict adaptation to directions orthogonal to a retained historical feature subspace to mitigate interference. However, historical and current tasks may rely on shared directions: excluding them can discard useful adaptation capacity, while low historical response alone does not ensure relevance to the incoming task. We therefore treat historical response as a graded metric for subspace selection. A previously active direction remains eligible when its current-task response is sufficiently strong relative to its historical exposure. Our guiding principle is: *select directions by their current-to-history response, then balance updates within the selected subspace.*

Relative-Response Spectral Allocation (RRS). Let $X_t \in \mathbb{R}^{(n_t n_b)\times m}$ collect the input features of task t, where n_t is the number of samples and n_b is the number of patch tokens per sample. Let $M_t =$

$(X_t)^\top X_t$ denote the current-task second-moment statistics and $M_{1:t-1} = \sum_{i=1}^{t-1} M_i$ the accumulated historical statistics. We define the regularized historical metric as $H_{t-1} = M_{1:t-1} + \epsilon I$, where $\epsilon > 0$. For a direction $v \neq 0$, we measure its relative response by $\mathcal{R}_t(v) = \frac{v^\top M_t v}{v^\top H_{t-1} v} = \frac{\|X_t v\|_2^2}{\sum_{i<t} \|X_i v\|_2^2 + \epsilon \|v\|_2^2}$. As illustrated in Figure 4, this criterion favors directions with strong current-task response relative to their historical response. Directions shared with previous tasks remain available when they are sufficiently responsive to the current task. To select a rank-r subspace, we solve

$$\max_{Z \in \mathbb{R}^{m \times r}} \operatorname{tr}(Z^\top M_t Z) \quad \text{s.t.} \quad Z^\top H_{t-1} Z = I_r. \tag{1}$$

The constraint gives each basis direction the same regularized historical-response budget and makes the directions orthogonal under the historical metric. Within this budget, the objective selects a subspace with the largest total current-task response. This provides a feature-based criterion for allocating adaptation capacity, without directly estimating task loss or forgetting.

Solving the subspace selection problem with whitening. We solve Eq. (1) by transforming the historical-metric constraint into a standard orthogonality constraint. Let $H_{t-1} = L_{t-1} L_{t-1}^\top$ be the Cholesky factorization of $H_{t-1} \succ 0$. With the change of variables $U = L_{t-1}^\top Z$, or equivalently $Z = L_{t-1}^{-\top} U$, Eq. (1) becomes

$$\max_{U^\top U = I_r} \operatorname{tr}\left(U^\top \widetilde{M}_t U\right), \widetilde{M}_t = L_{t-1}^{-1} M^t L_{t-1}^{-\top}.$$

Hence, the optimal whitened subspace is spanned by the leading r eigenvectors of $\widetilde{M}_t$, which correspond to directions with the strongest current-task response relative to historical response.

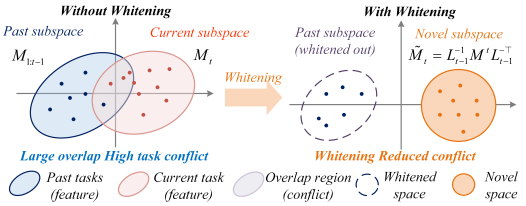


Figure 4: Allocating LoRA capacity using current-task response relative to historical feature statistics.

In practice, we approximate this dominant eigenspace using randomized SVD Halko et al. (2011): $\widetilde{U}_t, \Sigma_t, \widetilde{V}_t = \texttt{RSVD}\left(\widetilde{M}_t, r\right)$, where $\widetilde{U}_t \in \mathbb{R}^{m \times r}$ contains orthonormal approximate leading spectral directions in the whitened coordinates. We then map them back to the original feature space as $Z_t = L_{t-1}^{-\top} \widetilde{U}_t$. Because the mapped basis Z_t is generally not Euclidean-orthonormal, we perform a thin QR factorization $Z_t = Q_t R_t, Q_t^\top Q_t = I_r$, and define the fixed LoRA down-projection as $A_t = Q_t^\top$. Since R_t is nonsingular whenever Z_t has full column rank, $\operatorname{range}(Q_t) = \operatorname{range}(Z_t)$, so the QR step changes only the basis representation and preserves the selected relative-response subspace exactly. Meanwhile, $A_t A_t^\top = I_r$, which ensures that every factor update δB_t and its induced weight update $\delta B_t A_t$ have identical nonzero singular values. Thus, RRS preserves the generalized-eigen subspace selected by the relative-response criterion while providing the Euclidean-orthonormal parameterization required for exact spectral transfer in the subsequent SSB optimization. For the first task, no historical statistics are available and $H_0 = \epsilon I$. Therefore, $\widetilde{M}_1 = \epsilon^{-1} M_1$, and RRS reduces to selecting the dominant current-task feature subspace.

3.2 Subspace Spectral Balancing

Motivation: spectral collapse limits capacity utilization. Even within a task-responsive subspace, adaptation updates can concentrate most of their energy in a few singular modes, leaving the remaining directions weakly utilized. We refer to this concentration as *spectral collapse*: the effective rank of the update can be much lower than the nominal LoRA rank, even without a loss of algebraic rank. Subspace selection alone therefore does not ensure that the available capacity is effectively used. To address this issue, we apply Muon-style spectral balancing to the optimization step, reducing the dominance of a few modes within the selected subspace.

Connecting factor steps to weight steps. For task t, we parameterize the adapted weight as $W_{t,k} = W_{t-1} + B_{t,k} A_t, A_t A_t^\top = I_r$, where k indexes optimization steps. We freeze A_t and train $B_{t,k}$ from $B_{t,0} = 0$. Unlike updating both factors, this yields a weight step $S_k = \delta B_k A_t$, where $\delta B_k = B_{t,k+1} - B_{t,k}$, satisfying $S_k S_k^\top = \delta B_k A_t A_t^\top \delta B_k^\top = \delta B_k \delta B_k^\top$. Thus, the factor step and the weight step have identical nonzero singular values. Controlling the spectrum of δB_k therefore controls the spectrum of the actual weight step, without distortion from A_t.

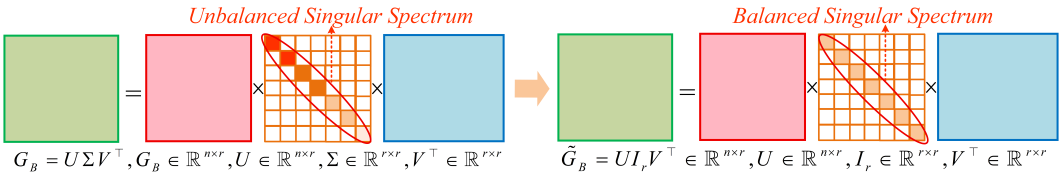


Figure 5: Spectral control of LoRA updates. An orthonormal input basis transfers the singular spectrum of each factor update to the induced weight update without distortion.

Subspace Spectral Balancing (SSB). Let $G_k \in \mathbb{R}^{n\times r}$ denote the momentum matrix used to update $B_{t,k}$. For $G_k \neq 0$, write its compact SVD as $G_k = U_k\Sigma_k V_k^\top, \Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_q)$, where $q = \mathrm{rank}(G_k)$ and $\sigma_j > 0$. The ideal polar transformation preserves the active singular directions and sets their singular values to one:

$$D_k = \mathcal{O}(G_k) = U_k V_k^\top, \; B_{t,k+1} = B_{t,k} - \eta_k \mathcal{O}(G_k). \tag{2}$$

This distributes the step energy equally across all active modes in Figure 5. SSB thus maximizes spectral participation on the existing update support; it does not create additional modes when G_k is rank deficient. Together, RRS and SSB determine which input directions are available for adaptation and how evenly each step uses its active singular modes.

Efficient approximate implementation. Computing an SVD at every step is unnecessary. Following Muon Jordan et al. (2024); Bernstein (2025), we approximate the polar transformation using Newton–Schulz polynomial iterations. Starting from $Y_0 = G_k/(\|G_k\|_F + \delta)$, with $\delta > 0$, we compute

$$K_j = Y_j^\top Y_j, \; Y_{j+1} = aY_j + bY_jK_j + cY_jK_j^2, \tag{3}$$

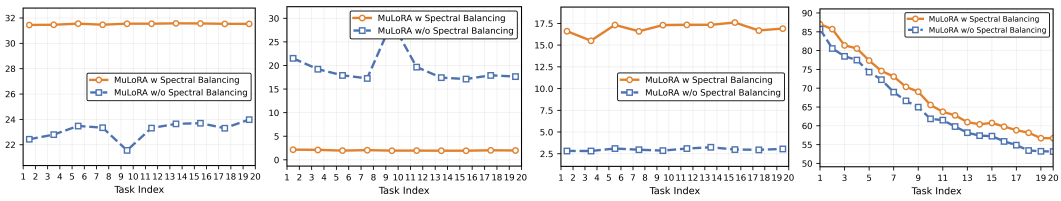
for $j = 0, \ldots, J-1$, and update $B_{t,k+1} = B_{t,k} - \eta_k Y_J$. Each iteration reshapes the singular values through a low-degree polynomial. Since $K_j \in \mathbb{R}^{r\times r}$, the computation exploits the small LoRA dimension. Finite iterations provide approximate spectral balancing, while spectral transfer from the factor step to the induced weight step remains exact. The same spectrum-preserving identity holds between $B_{t,k}$ and the accumulated adapter $B_{t,k}A_t$. However, balanced individual steps need not yield a balanced sum, because successive steps can reinforce or cancel different directions. Our theoretical analysis characterizes cumulative spectral behavior under controlled cross-step anisotropy.

4 Theoretical Analysis

Subspace-wide relative-response optimality. RRS first identifies a dominant subspace in the historically whitened coordinates and then maps it back to the original feature space. The following result shows that this mapped subspace is max–min optimal for the original relative-response criterion, while Euclidean re-orthogonalization preserves the guarantee.

Proposition 4.1 (Subspace-Wide Relative-Response Optimality). *Let $H_{t-1} = L_{t-1}L_{t-1}^\top \succ 0$ and $M_t \succeq 0$, and define $\widetilde{M}_t = L_{t-1}^{-1}M^tL_{t-1}^{-\top}$. Let $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ be the eigenvalues of $\widetilde{M}_t$, and let $U_r \in \mathbb{R}^{m\times r}$ contain its leading r orthonormal eigenvectors. Define the corresponding subspace in the original feature space as $\mathcal{S}_t = \mathrm{range}\big(L_{t-1}^{-\top}U_r\big)$. Then every nonzero $v \in \mathcal{S}_t$ satisfies $\lambda_r \leq \frac{v^\top M^t v}{v^\top H_{t-1} v} \leq \lambda_1$. Moreover, $\max_{\dim(\mathcal{S})=r} \min_{v\in\mathcal{S}\setminus\{0\}} \frac{v^\top M^t v}{v^\top H_{t-1} v} = \lambda_r$, and the maximum is attained by $\mathcal{S}_t$. If $L_{t-1}^{-\top}U_r = Q_tR_t$ is a thin QR factorization and $A_t = Q_t^\top$, then $A_tA_t^\top = I_r, \mathrm{row}(A_t) = \mathcal{S}_t$. Hence, Euclidean re-orthogonalization preserves the max–min-optimal RRS subspace while providing the row-orthonormal basis required for exact spectral transfer.*

Proposition 4.1 provides a uniform guarantee over the entire LoRA adaptation space: even its least responsive direction achieves relative response at least λ_r, and no other r-dimensional subspace can attain a larger worst-direction response. The result follows from the change of variables $u = L_{t-1}^\top v$, under which $\frac{v^\top M_t v}{v^\top H_{t-1} v} = \frac{u^\top \widetilde{M}_t u}{u^\top u}$, together with the Courant–Fischer theorem. Importantly, the subsequent QR factorization changes only the basis of the selected subspace, not the subspace itself. Thus, RRS simultaneously preserves its relative-response optimality and yields $A_tA_t^\top = I_r$, directly connecting subspace allocation to the spectrum-preserving LoRA parameterization used by SSB.



(a) Effective rank **(b)** κ_+ **(c)** Stable rank **(d)** Overall Test Acc

Figure 6: **Spectral balancing prevents rank collapse.** We track effective rank, κ_+, stable rank of ΔW, and overall test accuracy across tasks, comparing MuLoRA with and without spectral balancing. SSB maintains high effective rank and stable rank while keeping the nonzero spectrum well conditioned throughout sequential adaptation. Importantly, this geometric advantage persists over the same tasks for which overall class-incremental accuracy is better preserved, suggesting that spectral utilization is closely associated with retained adaptation capacity.

4.1 Cumulative Spectral Balance

Stepwise orthogonalization balances individual updates, but their accumulation also depends on cross-step interactions. We characterize how cumulative anisotropy controls the condition number, stable rank, and entropy-based effective rank of the accumulated update. For a nonzero matrix B with positive singular values $\sigma_1(B), \dots, \sigma_q(B)$, define $\mathrm{sr}(B) = \frac{\|B\|_F^2}{\|B\|_2^2}, \mathrm{erank}(B) = \exp\left(-\sum_{j=1}^q p_j \log p_j\right), p_j = \frac{\sigma_j(B)}{\sum_{i=1}^q \sigma_i(B)}$. Here $\log$ denotes the natural logarithm. For a full-column-rank matrix $B \in \mathbb{R}^{n\times r}$, let $\kappa(B) = \sigma_1(B)/\sigma_r(B)$. For a general nonzero matrix, κ_+ denotes the ratio of its largest to smallest positive singular value in Figure 6.

Theorem 4.2 (Cumulative Spectral Balance)**.** *Let $K \geq 1$, $n \geq r \geq 1$, $B_0 = 0$, and $B_K = -\sum_{k=1}^K \eta_k D_k, \eta_k > 0, D_k \in \mathbb{R}^{n\times r}, D_k^\top D_k = I_r$. Define $C_K = \sum_{1\leq i<j\leq K} \eta_i\eta_j \left(D_i^\top D_j + D_j^\top D_i\right)$, $a_K = \sum_{k=1}^K \eta_k^2 + \frac{\mathrm{tr}(C_K)}{r}, E_K = C_K - \frac{\mathrm{tr}(C_K)}{r} I_r$. Suppose $a_K > 0$ and $\|E_K\|_2 \leq \rho a_K$ for some $\rho \in [0,1)$. Then B_K has full column rank, and $\sqrt{a_K(1-\rho)} \leq \sigma_j(B_K) \leq \sqrt{a_K(1+\rho)}, j = 1, \dots, r$. Moreover, $\kappa(B_K) = \sqrt{\frac{a_K+\lambda_{\max}(E_K)}{a_K+\lambda_{\min}(E_K)}} \leq \sqrt{\frac{1+\rho}{1-\rho}}, \mathrm{sr}(B_K) = \frac{ra_K}{a_K+\lambda_{\max}(E_K)} \geq \frac{r}{1+\rho}, \mathrm{erank}(B_K) \geq \frac{r}{1+\|E_K\|_F^2/(ra_K^2)} \geq \frac{r}{1+\rho^2}$. If $A \in \mathbb{R}^{r\times m}$ satisfies $AA^\top = I_r$, then $\kappa_+(B_K A) = \kappa(B_K), \mathrm{sr}(B_K A) = \mathrm{sr}(B_K), \mathrm{erank}(B_K A) = \mathrm{erank}(B_K)$.*

Theorem 4.2 follows from the decomposition $B_K^\top B_K = a_K I_r + E_K$. The scalar $a_K = \|B_K\|_F^2/r$ determines the average squared singular value, whereas E_K/a_K measures relative spectral anisotropy. A smaller admissible ρ tightens the condition-number upper bound and raises the stable-rank and effective-rank lower bounds. When $E_K = 0$, all singular values are equal, giving $\kappa(B_K) = 1$ and $\mathrm{sr}(B_K) = \mathrm{erank}(B_K) = r$. These guarantees are conditional on cumulative anisotropy: stepwise orthogonalization alone does not ensure that the condition holds.

Corollary 4.3 (Robust Cumulative Spectral Balance)**.** *Let $K \geq 1$, $n \geq r \geq 1$, and $B_K = -\sum_{k=1}^K \eta_k D_k, \eta_k > 0, D_k \in \mathbb{R}^{n\times r}$, where $\|D_k^\top D_k - I_r\|_2 \leq \varepsilon_k, \varepsilon_k \geq 0$. Define C_K, E_K, and a_K as in Theorem 4.2, and let $\delta_K = \sum_{k=1}^K \eta_k^2 \varepsilon_k$. Suppose $a_K > 0$ and $\|E_K\|_2 \leq \rho a_K$ for some $\rho \in [0,1)$. Then $a_K(1-\rho) - \delta_K \leq \sigma_j(B_K)^2 \leq a_K(1+\rho) + \delta_K, j = 1, \dots, r$. If $\delta_K < a_K(1-\rho)$, define $\tau_K = \delta_K/a_K$. Then B_K has full column rank, and $\kappa(B_K) \leq \sqrt{\frac{1+\rho+\tau_K}{1-\rho-\tau_K}}, \mathrm{sr}(B_K) \geq \frac{r(1-\tau_K)}{1+\rho+\tau_K}, \mathrm{erank}(B_K) \geq \frac{r}{1+\left(\frac{\rho+\tau_K}{1-\tau_K}\right)^2}$. If $A \in \mathbb{R}^{r\times m}$ satisfies $AA^\top = I_r$, the same bounds hold for $\kappa_+(B_K A)$, $\mathrm{sr}(B_K A)$, and $\mathrm{erank}(B_K A)$, respectively.*

Corollary 4.3 separates cross-step anisotropy ρ from the relative accumulated orthogonalization error τ_K. The condition $\rho + \tau_K < 1$ ensures a positive lower bound on the smallest squared singular value. When $\tau_K = 0$, the bounds reduce to those in Theorem 4.2. Proofs are provided in the appendix.

Table 1: Experimental results for 10 incremental tasks on ImageNetR, ImageNetA, CIFAR100, and CUB. We report the averaged results over 3 random seeds. Best and second-best results are **bold** and underlined.

Method	Venue	10S-ImageNetR		10S-ImageNetA		10S-CIFAR100		10S-CUB	
		$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$
L2P	CVPR'22	$72.34_{\pm 0.17}$	$77.36_{\pm 0.64}$	$44.04_{\pm 0.93}$	$51.24_{\pm 2.26}$	$84.06_{\pm 0.88}$	$88.26_{\pm 1.34}$	$67.02_{\pm 1.90}$	$79.62_{\pm 1.60}$
DualPrompt	ECCV'22	$69.10_{\pm 0.62}$	$74.28_{\pm 0.66}$	$53.19_{\pm 0.74}$	$64.59_{\pm 0.08}$	$86.93_{\pm 0.24}$	$91.13_{\pm 0.32}$	$68.48_{\pm 0.47}$	$80.59_{\pm 1.50}$
CODAPrompt	CVPR'23	$73.31_{\pm 0.50}$	$78.47_{\pm 0.53}$	$52.08_{\pm 0.12}$	$63.92_{\pm 0.12}$	$83.21_{\pm 3.39}$	$87.71_{\pm 3.17}$	$77.23_{\pm 1.12}$	$81.90_{\pm 0.85}$
InfLoRA	CVPR'24	$74.75_{\pm 0.64}$	$80.67_{\pm 0.55}$	$49.20_{\pm 1.12}$	$60.92_{\pm 0.61}$	$86.75_{\pm 0.35}$	$91.72_{\pm 0.15}$	$70.82_{\pm 0.23}$	$81.39_{\pm 0.14}$
SD-LoRA	ICLR'25	$77.34_{\pm 0.35}$	$82.04_{\pm 0.24}$	$55.96_{\pm 0.73}$	$64.95_{\pm 1.63}$	$88.01_{\pm 0.31}$	$92.54_{\pm 0.18}$	$77.48_{\pm 0.20}$	$85.59_{\pm 0.44}$
Bi-LoRA	CVPR'25	$77.95_{\pm 0.14}$	$81.52_{\pm 0.26}$	–	–	$87.46_{\pm 0.76}$	$92.50_{\pm 0.62}$	–	–
PLAN	ICCV'25	$75.25_{\pm 0.42}$	$80.41_{\pm 0.56}$	–	–	$87.54_{\pm 0.31}$	$92.21_{\pm 0.35}$	–	–
LoRA-P&M	NeurIPS'25	$79.95_{\pm 0.18}$	$85.29_{\pm 0.93}$	$56.57_{\pm 0.78}$	$65.35_{\pm 1.81}$	$88.45_{\pm 0.35}$	$92.89_{\pm 1.13}$	$78.29_{\pm 0.50}$	$83.39_{\pm 0.61}$
CoSO	NeurIPS'25	$\underline{81.10}_{\pm 0.39}$	$85.56_{\pm 0.13}$	–	–	$88.77_{\pm 0.16}$	$92.99_{\pm 0.23}$	–	–
SplitLoRA	ICLR'26	$81.00_{\pm 0.17}$	$\underline{85.84}_{\pm 0.62}$	–	–	$\mathbf{90.33}_{\pm \mathbf{0.73}}$	$\underline{93.70}_{\pm 0.32}$	–	–
EBLoRA	ICML'26	$79.56_{\pm 0.68}$	$82.68_{\pm 0.86}$	$\underline{61.56}_{\pm 0.54}$	$\underline{69.28}_{\pm 0.56}$	$88.98_{\pm 0.61}$	$92.86_{\pm 0.42}$	$\underline{78.92}_{\pm 0.38}$	$\underline{88.64}_{\pm 0.52}$
EWC-LoRA	ICLR'26	$72.86_{\pm 0.79}$	$78.95_{\pm 0.86}$	$59.89_{\pm 0.26}$	$68.33_{\pm 0.67}$	$87.91_{\pm 0.57}$	$92.27_{\pm 0.39}$	$72.62_{\pm 0.32}$	$83.52_{\pm 0.54}$
MuLoRA (Ours)	–	$\mathbf{81.77}_{\pm \mathbf{0.25}}$	$\mathbf{87.28}_{\pm \mathbf{0.28}}$	$\mathbf{62.23}_{\pm \mathbf{0.52}}$	$\mathbf{71.27}_{\pm \mathbf{1.02}}$	$\underline{90.03}_{\pm 0.26}$	$\mathbf{93.81}_{\pm \mathbf{1.42}}$	$\mathbf{80.27}_{\pm \mathbf{0.25}}$	$\mathbf{89.47}_{\pm \mathbf{0.42}}$

Table 2: Experimental results for 5 and 20 incremental sessions on ImageNetR, 20 sessions on ImageNetA, and 10 sessions on DomainNet. The highest results are in **bold**, and the second highest results are underlined.

Method	Venue	5S-ImageNetR		20S-ImageNetR		20S-ImageNetA		10S-DomainNet	
		$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$
L2P	CVPR'22	$70.83_{\pm 0.58}$	$78.34_{\pm 0.47}$	$69.64_{\pm 0.42}$	$75.28_{\pm 0.57}$	$40.48_{\pm 1.78}$	$49.62_{\pm 1.46}$	$81.17_{\pm 0.83}$	$87.43_{\pm 0.95}$
DualPrompt	ECCV'22	$73.05_{\pm 0.50}$	$79.47_{\pm 0.40}$	$66.61_{\pm 0.58}$	$72.45_{\pm 0.37}$	$42.28_{\pm 1.94}$	$53.39_{\pm 1.64}$	$81.70_{\pm 0.78}$	$87.80_{\pm 0.99}$
CODAPrompt	CVPR'23	$74.91_{\pm 0.33}$	$79.25_{\pm 0.53}$	$69.96_{\pm 0.50}$	$75.34_{\pm 0.85}$	$44.62_{\pm 1.92}$	$54.86_{\pm 0.50}$	$80.04_{\pm 0.79}$	$86.27_{\pm 0.82}$
InfLoRA	CVPR'24	$77.52_{\pm 0.37}$	$82.01_{\pm 0.12}$	$71.01_{\pm 0.45}$	$77.28_{\pm 0.45}$	$46.81_{\pm 2.30}$	$57.13_{\pm 0.95}$	$81.45_{\pm 0.68}$	$87.55_{\pm 0.57}$
SD-LoRA	ICLR'25	$79.15_{\pm 0.20}$	$83.01_{\pm 0.42}$	$75.26_{\pm 0.37}$	$80.22_{\pm 0.72}$	$49.12_{\pm 1.78}$	$59.63_{\pm 1.57}$	$81.01_{\pm 0.42}$	$86.85_{\pm 0.40}$
Bi-LoRA	CVPR'25	$78.32_{\pm 0.37}$	$83.31_{\pm 0.52}$	$72.41_{\pm 0.65}$	$79.28_{\pm 0.76}$	–	–	$76.56_{\pm 0.41}$	$83.20_{\pm 0.35}$
LoRA-P&M	NeurIPS'25	$81.47_{\pm 0.56}$	$85.96_{\pm 0.52}$	$76.37_{\pm 0.09}$	$82.77_{\pm 0.31}$	$52.27_{\pm 1.84}$	$62.15_{\pm 1.33}$	$\underline{84.35}_{\pm 0.48}$	$\underline{89.43}_{\pm 0.27}$
CoSO	NeurIPS'25	$82.10_{\pm 0.13}$	$\underline{86.38}_{\pm 0.07}$	$78.19_{\pm 0.28}$	$83.69_{\pm 0.12}$	–	–	–	–
SplitLoRA	ICLR'26	$81.92_{\pm 0.29}$	$85.85_{\pm 0.55}$	$\underline{78.82}_{\pm 0.28}$	$\underline{84.57}_{\pm 0.44}$	–	–	$84.31_{\pm 0.23}$	$88.99_{\pm 0.57}$
EBLoRA	ICML'26	$\underline{82.56}_{\pm 0.32}$	$86.11_{\pm 0.25}$	$77.45_{\pm 0.23}$	$83.76_{\pm 0.65}$	$\underline{52.52}_{\pm 1.01}$	$\underline{62.16}_{\pm 0.82}$	$83.67_{\pm 0.14}$	$89.27_{\pm 0.21}$
EWC-LoRA	ICLR'26	$76.36_{\pm 0.25}$	$82.12_{\pm 0.28}$	$70.18_{\pm 0.15}$	$78.56_{\pm 0.74}$	$48.58_{\pm 1.15}$	$59.06_{\pm 0.82}$	$83.56_{\pm 0.28}$	$89.25_{\pm 0.12}$
MuLoRA (Ours)	–	$\mathbf{83.21}_{\pm 0.24}$	$\mathbf{87.56}_{\pm 0.32}$	$\mathbf{78.93}_{\pm 0.18}$	$\mathbf{85.02}_{\pm 0.21}$	$\mathbf{55.86}_{\pm 0.61}$	$\mathbf{65.61}_{\pm 0.64}$	$\mathbf{85.61}_{\pm 0.10}$	$\mathbf{90.21}_{\pm 0.32}$

5 EXPERIMENTS

Datasets and Evaluation Metrics. We evaluate our method on five datasets, namely ImageNetR Hendrycks et al. (2021a), ImageNetA Hendrycks et al. (2021b), CIFAR-100 Krizhevsky et al. (2009), CUB Wah et al. (2011) and DomainNet Peng et al. (2019). ImageNet-R contains 30,000 images from 200 categories and diverse styles. ImageNet-A includes 7,500 real-world adversarial examples, spanning 200 classes. CIFAR-100 is a standard CL benchmark, including 60,000 32×32 images from 100 classes. DomainNet consists of images from 345 classes across 6 visual domains. Following prior works Gao et al. (2024), we select the top 200 most populous classes for evaluation. Following existing PEFT-based methods Tan et al. (2024); Wu et al. (2025a), we report results using 2 metrics: (1) Last accuracy (A_{last}), defined as the accuracy over all classes after the final session and (2) Average accuracy (A_{Avg}), defined as the mean accuracy across all incremental sessions.

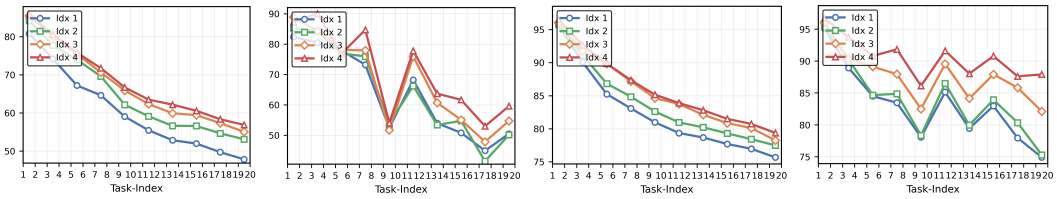
Implementation Details. All experiments are conducted on a single NVIDIA RTX 4090 GPU. We use ViT-B/16 Dosovitskiy et al. (2021) pre-trained on ImageNet-21K Russakovsky et al. (2015) as the backbone. The model is optimized with the Muon Jordan et al. (2024) with a Cosine Annealing scheduler. We train each task for 20 epochs with a batch size of 48. The initial learning rate is set to 0.0001 for ImageNetR, and 0.0002 for others. Following Wu et al. (2025a), we insert MuLoRA into all attention blocks with rank 32, and employ a cosine classifier for the classification task.

5.1 COMPARISON WITH BASELINE METHODS

We compare MuLoRA with representative prompt-based, adapter-based, and LoRA-based continual learning methods Wang et al. (2022b;a); Jiao et al. (2024); Tan et al. (2024); Gao et al. (2023); Liang & Li (2024); Wu et al. (2025b); Wang et al. (2025); Cheng et al. (2025); Qiu et al. (2025b). As shown in Tables 1 and 2, MuLoRA achieves the highest mean accuracy in 15 of 16 reported metrics across five benchmarks and eight incremental settings. These results support jointly addressing low-rank capacity allocation and spectral utilization. On 10S-ImageNetR and 10S-ImageNetA, MuLoRA achieves 81.77% and 62.23% A_{Last}, respectively, exceeding the strongest reported baseline for each setting by 0.67 percentage points. The corresponding A_{Avg} gains are 1.44 and 1.99 points, indicating

Table 3: Ablation studies on ImageNetR and ImageNetA under 10 and 20 incremental sessions, evaluating the individual and combined contributions of Relative-Response Spectral Allocation (RRS) and Subspace Spectral Balancing (SSB). Idx 1 uses a fixed random orthonormal A and trains only B.

Idx	RRS	SSB	10S-ImageNetR		10S-ImageNetA		20S-ImageNetR		20S-ImageNetA	
			$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$
1	–	–	$76.23_{\pm 0.18}$	$82.05_{\pm 0.62}$	$57.48_{\pm 1.47}$	$67.34_{\pm 1.86}$	$76.21_{\pm 0.18}$	$80.26_{\pm 0.51}$	$48.62_{\pm 1.28}$	$60.52_{\pm 1.26}$
2	✓	–	$79.23_{\pm 0.28}$	$85.34_{\pm 0.36}$	$59.77_{\pm 0.56}$	$68.91_{\pm 1.08}$	$77.86_{\pm 0.20}$	$83.06_{\pm 0.32}$	$51.92_{\pm 0.66}$	$62.42_{\pm 0.51}$
3	–	✓	$80.74_{\pm 0.28}$	$86.31_{\pm 0.35}$	$61.86_{\pm 0.55}$	$70.26_{\pm 0.98}$	$78.50_{\pm 0.26}$	$84.25_{\pm 0.36}$	$54.32_{\pm 0.64}$	$64.53_{\pm 0.51}$
4	✓	✓	$\mathbf{81.77}_{\pm 0.25}$	$\mathbf{87.28}_{\pm 0.28}$	$\mathbf{62.23}_{\pm 0.52}$	$\mathbf{71.27}_{\pm 1.02}$	$\mathbf{78.93}_{\pm 0.18}$	$\mathbf{85.02}_{\pm 0.21}$	$\mathbf{55.86}_{\pm 0.61}$	$\mathbf{65.61}_{\pm 0.64}$



(a) 20S-ImageNetA (all) **(b)** 20S-ImageNetA (new) **(c)** 20S-ImageNetR (all) **(d)** 20S-ImageNetR (new)

Figure 7: Stability–plasticity dynamics across incremental tasks. Overall and new-task accuracy on 20S-ImageNetA and 20S-ImageNetR. MuLoRA better maintains overall performance while preserving strong adaptation to newly introduced classes.

improvements across the learning sequence as well as at its endpoint. The largest gains occur on 20S-ImageNetA: MuLoRA reaches 55.86% A_{Last} and 65.61% A_{Avg}, improving over LoRA-P&M by 3.59 and 3.46 points, respectively. On ImageNetR, it also obtains the highest means under both 5- and 20-session protocols, although the 20-session A_{Last} margin is modest (0.11 points). The benefits extend to CUB and DomainNet, where MuLoRA improves A_{Last} over the strongest reported baselines by 1.35 and 1.26 points, respectively. On CIFAR-100, it achieves the highest A_{Avg} (93.81%), while its A_{Last} (90.03%) is 0.30 points below SplitLoRA. Overall, the results demonstrate broad empirical benefits from combining task-responsive subspace selection with spectrally controlled updates, without replaying past examples or historical features.

5.2 Ablation Studies

Effectiveness of RRS and SSB. Table 3 and Figure 7 evaluate the individual and combined effects of RRS and SSB on ImageNetR and ImageNetA under 10- and 20-session protocols. Both components improve performance across all four settings. RRS alone increases A_{Last} by 1.65–3.30 percentage points over the baseline, supporting the benefit of selecting adaptation directions based on current-task response relative to historical exposure. SSB alone improves A_{Last} by 2.29–5.70 points, yielding larger gains than RRS alone in each setting and supporting the effectiveness of spectral balancing within the adaptation space. Combining RRS and SSB achieves the highest mean accuracy on all eight reported metrics. Compared with SSB alone, adding RRS further improves A_{Last} by 0.37–1.54 points and A_{Avg} by 0.77–1.08 points. On 20S-ImageNetA, the full model reaches 55.86% A_{Last} and 65.61% A_{Avg}, exceeding the baseline by 7.24 and 5.09 points, respectively. These results support the complementary benefits of task-responsive subspace selection and spectral balancing: their combination consistently outperforms either component alone.

6 Conclusion

We introduced MuLoRA to address ***spectral plasticity collapse*** that can leave low-rank adaptation capacity underutilized in continual learning. MuLoRA couples history-aware subspace selection with Muon-style spectral balancing through a fixed row-orthonormal basis, transferring factor-step spectra exactly to weight space. Our analysis establishes the max–min optimality of exact subspace selection and provides conditional spectral guarantees for accumulated updates. Experiments across five benchmarks and eight incremental settings support a central design principle: effective adaptation requires both selecting task-responsive directions and controlling how updates use them.

REFERENCES

Jeremy Bernstein. Deriving Muon, 2025. URL https://jeremybernste.in/writing/deriving-muon. Technical blog post.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23595–23604, 2024.

Quan Cheng, Yuanyu Wan, Lingyu Wu, Chenping Hou, and Lijun Zhang. Continuous subspace optimization for continual learning. In *Conference on Neural Information Processing Systems*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11483–11493, 2023.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28463–28473, 2024.

Hao Gu, Mao-Lin Luo, Zi-Hao Zhou, Han-Chen Zhang, Min-Ling Zhang, and Tong Wei. Spectral imbalance causes forgetting in low-rank continual adaptation. *arXiv preprint arXiv:2602.00722*, 2026. doi: 10.48550/arXiv.2602.00722. URL https://arxiv.org/abs/2602.00722.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Lingfeng He, De Cheng, Zhiheng Ma, Huaijie Wang, Dingwen Zhang, Nannan Wang, and Xinbo Gao. Ckaa: Cross-subspace knowledge alignment and aggregation for robust continual learning. *arXiv preprint arXiv:2507.09471*, 2025.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li Jiao, Qiuxia Lai, Yu Li, and Qiang Xu. Vector quantization prompting for continual learning. *NeurIPS*, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jiashuo Li, Shaokun Wang, Bo Qian, Yuhang He, Xing Wei, Qiang Wang, and Yihong Gong. Dynamic integration of task-specific adapters for class incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30545–30555, 2025.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Yue Lu, Shizhou Zhang, De Cheng, Yinghui Xing, Nannan Wang, Peng Wang, and Yanning Zhang. Visual Prompt Tuning in Null Space for Continual Learning. In *Conference on Neural Information Processing Systems*, 2024.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Jingyang Qiao, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, Yuan Xie, et al. Prompt gradient projection for continual learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Haomiao Qiu, Miao Zhang, Ziyue Qiao, Weili Guan, Min Zhang, and Liqiang Nie. Splitlora: Balancing stability and plasticity in continual learning through gradient space splitting. *arXiv preprint arXiv:2505.22370*, 2025a.

Haomiao Qiu, Miao Zhang, Ziyue Qiao, and Liqiang Nie. Train with perturbation, infer after merging: A two-stage framework for continual learning. In *Conference on Neural Information Processing Systems*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023.

Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. Semantically-shifted incremental adapter-tuning is a continual vitransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23252–23262, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Huaijie Wang, De Cheng, Lingfeng He, Yan Li, Jie Li, Nannan Wang, and Xinbo Gao. Ekpc: Elastic knowledge preservation and compensation for class-incremental learning: H. wang et al. *International Journal of Computer Vision*, 134(5):238, 2026.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023.

Xiequn Wang, Zhan Zhuang, and Yu Zhang. Plan: Proactive low-rank allocation for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2909–2918, 2025.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Fangwen Wu, Lechao Cheng, Shengeng Tang, Xiaofeng Zhu, Chaowei Fang, Dingwen Zhang, and Meng Wang. Navigating semantic drift in task-agnostic class-incremental learning. In *Forty-second International Conference on Machine Learning*, 2025a.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. SD-loRA: Scalable decoupled low-rank adaptation for class incremental learning. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=5U1rlpX68A.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido M. van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pZj2DhfaVD.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25613–25622, 2025.

ETHICS STATEMENT

This work studies parameter-efficient continual learning using existing image classification benchmarks: ImageNet-R, ImageNet-A, CIFAR-100, CUB, and DomainNet. Our experiments use benchmark datasets and pretrained models rather than collecting new data from human participants. Although MuLoRA aims to improve adaptation across sequential tasks, it does not provide formal guarantees of privacy, fairness, or robustness. Models may inherit biases from pretraining and downstream datasets, and aggregate classification accuracy does not establish reliable performance across demographic groups or deployment conditions. Deployment in sensitive applications therefore requires appropriate privacy safeguards, bias assessment, and application-specific evaluation.

REPRODUCIBILITY STATEMENT

The paper describes MuLoRA's task-adaptive subspace construction and momentum orthogonalization, together with the assumptions underlying its theoretical results. The appendix provides detailed proofs of the subspace selection characterization, conditional cumulative spectral bounds, and preservation of nonzero singular values under a row-orthonormal basis. The experimental section specifies the benchmark datasets, incremental settings, evaluation metrics, pretrained ViT-B/16 backbone, adapter placement and rank, optimizer, learning rates, learning-rate schedule, training epochs, and batch size. Experiments are conducted on a single NVIDIA RTX 4090 GPU, and the main comparison reports results averaged over three random seeds. We will publicly release the implementation and experiment scripts upon acceptance.

AI USE STATEMENT

We used large language models to assist with language editing, including sentence rephrasing, grammar correction, and readability improvements. We did not use these tools for research ideation, methodology or experiment design, theoretical derivations, method implementation, data generation or processing, or interpretation of results. The authors reviewed and revised all AI-assisted text and take full responsibility for the final manuscript, including its claims, analyses, and conclusions.

APPENDIX CONTENTS

A Proofs of the Theoretical Results **15**

A.1 Two spectral preliminaries 15

A.2 Exact relative-response subspace selection 16

A.3 Cumulative spectral balance with exact steps 17

A.4 Cumulative spectral balance with approximate steps 18

B Approximate Computation and Scope of the Guarantees **19**

B.1 Interpreting historical response 19

B.2 A computable characterization of an approximate RRS subspace 19

B.3 Interpretation and scope of the cumulative guarantees 20

B.4 Checking the cumulative conditions without pairwise step storage 20

C Mechanism Comparisons on 20S-ImageNet-A **21**

C.1 Definitions of the compared variants 21

C.2 Results and interpretation 21

D Reproducibility Details and Limitations **22**

D.1 Reported experimental configuration 22

D.2 Memory and computational accounting 22

D.3 Scope of empirical and theoretical conclusions 22

A PROOFS OF THE THEORETICAL RESULTS

We prove the exact relative-response guarantee and the cumulative spectral bounds stated in the main text. All matrices are real, and all logarithms are natural. Singular values are ordered nonincreasingly. For the cumulative analysis, $n \geq r$; all r singular values of $B \in \mathbb{R}^{n\times r}$ are retained, including zeros. The cumulative analysis concerns optimization steps within one task, during which the input basis is fixed.

A.1 TWO SPECTRAL PRELIMINARIES

Spectral quantities. For a nonzero matrix B with positive singular values $s_1 \geq \cdots \geq s_q > 0$, we use

$$\mathrm{sr}(B) = \frac{\sum_{j=1}^{q} s_j^2}{s_1^2}, \qquad \kappa_+(B) = \frac{s_1}{s_q}, \qquad \mathrm{erank}(B) = \exp\left(-\sum_{j=1}^{q} p_j \log p_j\right), \quad p_j = \frac{s_j}{\sum_i s_i}. \tag{4}$$

Entropy-based effective rank differs from the energy participation ratio $(\sum_j s_j^2)^2/\sum_j s_j^4$; the next lemma relates them by inequalities. Neither is the algebraic rank. In particular, κ_+ excludes zero singular values and must be interpreted together with rank information. For the zero matrix, normalized spectral statistics are undefined.

Lemma A.1 (Preservation of the nonzero singular spectrum)**.** *Let $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$ satisfy $AA^\top = I_r$. Then BA and B have the same nonzero singular values, including multiplicities, and the same rank. Consequently, for $B \neq 0$, their stable ranks, entropy-based effective ranks, and nonzero condition numbers coincide. If B has full column rank, then $\kappa_+(BA) = \kappa(B)$.*

Proof. The left Gram matrices coincide:

$$(BA)(BA)^\top = BAA^\top B^\top = BB^\top. \tag{5}$$

The positive eigenvalues of a left Gram matrix are precisely the squares of the nonzero singular values of the corresponding matrix. Thus the nonzero singular spectra, their multiplicities, and the ranks agree. Each stated spectral statistic is determined entirely by these nonzero singular values, which proves the remaining claims. □

Lemma A.2 (Entropy effective rank and energy participation)**.** *Let $B \neq 0$ have positive singular values $s_1, \ldots, s_q$, and set*

$$S_\ell = \sum_{j=1}^{q} s_j^\ell, \qquad p_j = \frac{s_j}{S_1}, \qquad \mathrm{erank}(B) = \exp\left(-\sum_{j=1}^{q} p_j \log p_j\right). \tag{6}$$

Then

$$\mathrm{erank}(B) \geq \frac{S_1^2}{S_2} \geq \frac{S_2^2}{S_4} = \frac{\left(\mathrm{tr}(B^\top B)\right)^2}{\mathrm{tr}\left((B^\top B)^2\right)}. \tag{7}$$

The last quantity is the participation rank of the squared singular values; it is generally distinct from $\mathrm{erank}(B)$. *If all positive singular values are equal, all quantities in equation 7 equal q.*

Proof. By concavity of the logarithm, using the positive weights p_j,

$$\sum_{j=1}^{q} p_j \log p_j \leq \log\left(\sum_{j=1}^{q} p_j^2\right). \tag{8}$$

Negating and exponentiating gives

$$\mathrm{erank}(B) \geq \frac{1}{\sum_j p_j^2} = \frac{S_1^2}{S_2}. \tag{9}$$

Next, Hölder's inequality with conjugate exponents $3/2$ and 3 yields

$$S_2 = \sum_{j=1}^{q} s_j^{2/3} s_j^{4/3} \le \left(\sum_{j=1}^{q} s_j \right)^{2/3} \left(\sum_{j=1}^{q} s_j^4 \right)^{1/3} = S_1^{2/3} S_4^{1/3}. \tag{10}$$

Therefore $S_2^3 \le S_1^2 S_4$. Since $S_2, S_4 > 0$, division gives $S_1^2/S_2 \ge S_2^2/S_4$. Finally, $\mathrm{tr}(B^\top B) = S_2$ and $\mathrm{tr}((B^\top B)^2) = S_4$, including when B is rank deficient. Equal positive singular values give $p_j = 1/q$ and equality throughout. □

A.2 Exact relative-response subspace selection

This subsection proves Proposition 4.1.

Proof. Assume $1 \le r \le m$, and abbreviate $H = H_{t-1}$, $L = L_{t-1}$, $M = M^t$, and $\widetilde{M} = L^{-1} M L^{-\top}$. Since $H = LL^\top \succ 0$, L is invertible. Moreover, $\widetilde{M}$ is symmetric positive semidefinite.

Step 1: Transform the relative response. For every $v \neq 0$, set $u = L^\top v$, so that $v = L^{-\top} u$. Then

$$v^\top H v = u^\top u, \qquad v^\top M v = u^\top \widetilde{M} u. \tag{11}$$

Consequently,

$$\frac{v^\top M v}{v^\top H v} = \frac{u^\top \widetilde{M} u}{u^\top u}. \tag{12}$$

Thus historical whitening converts the generalized Rayleigh quotient into an ordinary Rayleigh quotient through an invertible change of coordinates.

Step 2: Bound every direction in the selected subspace. Let $\Lambda_r = \mathrm{diag}(\lambda_1, \dots, \lambda_r)$. Since $U_r^\top U_r = I_r$ and $\widetilde{M} U_r = U_r \Lambda_r$, every nonzero $v \in \mathcal{S}_t$ can be written uniquely as $v = L^{-\top} U_r c$ for some $c \in \mathbb{R}^r \setminus \{0\}$. Eq. equation 12 gives

$$\frac{v^\top M v}{v^\top H v} = \frac{c^\top \Lambda_r c}{c^\top c} = \frac{\sum_{j=1}^{r} \lambda_j c_j^2}{\sum_{j=1}^{r} c_j^2} \in [\lambda_r, \lambda_1]. \tag{13}$$

Choosing c to be the r-th coordinate vector attains the lower endpoint. Hence the minimum response over $\mathcal{S}_t \setminus \{0\}$ is exactly λ_r.

Step 3: Establish global max–min optimality. The map $\mathcal{S} \mapsto L^\top \mathcal{S}$ is a bijection on the set of r-dimensional subspaces. Therefore

$$\max_{\dim(\mathcal{S})=r} \min_{v \in \mathcal{S} \setminus \{0\}} \frac{v^\top M v}{v^\top H v} = \max_{\dim(\mathcal{U})=r} \min_{u \in \mathcal{U} \setminus \{0\}} \frac{u^\top \widetilde{M} u}{u^\top u} = \lambda_r. \tag{14}$$

For completeness, the last equality can also be verified directly. Extend the columns of U_r to an orthonormal eigenbasis $u_1, \dots, u_m$ of $\widetilde{M}$, ordered by eigenvalue. Every r-dimensional subspace $\mathcal{U}$ intersects $\mathrm{span}\{u_r, \dots, u_m\}$ nontrivially, since their dimensions sum to $m+1$. A nonzero vector in the intersection has Rayleigh quotient at most λ_r. Thus no such $\mathcal{U}$ has minimum Rayleigh quotient greater than λ_r. The subspace $\mathrm{range}(U_r)$ attains λ_r by equation 13. This is the relevant Courant–Fischer characterization.

Step 4: Verify preservation under Euclidean QR. The matrix $Z = L^{-\top} U_r$ has full column rank. In its thin QR factorization $Z = Q_t R_t$, $Q_t^\top Q_t = I_r$ and R_t is invertible. Hence

$$\mathrm{range}(Q_t) = \mathrm{range}(Z) = \mathcal{S}_t. \tag{15}$$

Defining $A_t = Q_t^\top$ therefore gives

$$A_t A_t^\top = I_r, \qquad \mathrm{row}(A_t) = \mathrm{range}(A_t^\top) = \mathcal{S}_t, \tag{16}$$

where row vectors are identified with their transposes. Thus QR preserves the max–min-optimal subspace and provides the orthonormal parameterization required by Lemma A.1. □

Scope of the subspace guarantee. The proposition concerns exact leading eigenvectors. Repeated eigenvalues may make the optimal subspace nonunique, but do not affect the result. With randomized SVD, QR still preserves the computed subspace exactly; global max–min optimality is not automatic for that approximate subspace. Also, QR need not preserve the constraint $Z^\top HZ = I_r$: it preserves the subspace and its relative-response guarantee, while replacing the basis by a Euclidean-orthonormal one.

A.3 Cumulative spectral balance with exact steps

This subsection proves Theorem 4.2. For clarity, the quantities used below are

$$C_K = \sum_{1\le i<j\le K} \eta_i\eta_j(D_i^\top D_j + D_j^\top D_i), \tag{17}$$

$$a_K = \sum_{k=1}^{K} \eta_k^2 + \frac{\operatorname{tr}(C_K)}{r}, \qquad E_K = C_K - \frac{\operatorname{tr}(C_K)}{r} I_r. \tag{18}$$

The assumptions are $B_0 = 0$, $\eta_k > 0$, $D_k^\top D_k = I_r$, $a_K > 0$, and $\|E_K\|_2 \le \rho a_K$ for $0 \le \rho < 1$.

Proof. For brevity, write $B = B_K$, $C = C_K$, $a = a_K$, and $E = E_K$.

Step 1: Expand the cumulative Gram matrix. Since $B = -\sum_{k=1}^{K} \eta_k D_k$ and $D_k^\top D_k = I_r$,

$$\begin{aligned} B^\top B &= \sum_{k=1}^{K} \eta_k^2 D_k^\top D_k + \sum_{1\le i<j\le K} \eta_i\eta_j(D_i^\top D_j + D_j^\top D_i) \\ &= \left(\sum_{k=1}^{K} \eta_k^2\right) I_r + C = aI_r + E. \end{aligned} \tag{19}$$

By definition E is symmetric and $\operatorname{tr}(E) = 0$. Taking traces gives

$$\|B\|_F^2 = \operatorname{tr}(B^\top B) = ra. \tag{20}$$

Step 2: Bound singular values and the condition number. Let $e_1 \ge \cdots \ge e_r$ be the eigenvalues of E. Eq. equation 19 implies $\sigma_j(B)^2 = a + e_j$. Because $|e_j| \le \|E\|_2 \le \rho a$,

$$a(1-\rho) \le \sigma_j(B)^2 \le a(1+\rho), \qquad j = 1, \dots, r. \tag{21}$$

The lower bound is strictly positive because $a > 0$ and $\rho < 1$. Thus B has full column rank, and taking square roots proves the singular-value bounds. In particular,

$$\kappa(B) = \sqrt{\frac{a + \lambda_{\max}(E)}{a + \lambda_{\min}(E)}} \le \sqrt{\frac{1+\rho}{1-\rho}}. \tag{22}$$

Step 3: Bound the stable rank. Using equation 20 and $\|B\|_2^2 = a + \lambda_{\max}(E)$ yields

$$\operatorname{sr}(B) = \frac{ra}{a + \lambda_{\max}(E)} \ge \frac{r}{1+\rho}. \tag{23}$$

Step 4: Bound the entropy-based effective rank. Since E is symmetric and traceless,

$$\begin{aligned} \operatorname{tr}\big((B^\top B)^2\big) &= \operatorname{tr}\big((aI_r + E)^2\big) \\ &= ra^2 + 2a\operatorname{tr}(E) + \operatorname{tr}(E^2) = ra^2 + \|E\|_F^2. \end{aligned} \tag{24}$$

Lemma A.2 therefore gives

$$\operatorname{erank}(B) \ge \frac{(ra)^2}{ra^2 + \|E\|_F^2} = \frac{r}{1 + \|E\|_F^2/(ra^2)}. \tag{25}$$

Finally, $\|E\|_F^2 \le r\|E\|_2^2 \le r\rho^2 a^2$, so

$$\operatorname{erank}(B) \ge \frac{r}{1+\rho^2}. \tag{26}$$

This argument bounds the entropy effective rank through energy participation; it does not equate the two definitions.

Step 5: Transfer the guarantees to the weight update. If $AA^\top = I_r$, Lemma A.1 implies

$$\kappa_+(BA) = \kappa(B), \qquad \mathrm{sr}(BA) = \mathrm{sr}(B), \qquad \mathrm{erank}(BA) = \mathrm{erank}(B). \tag{27}$$

Thus all stated bounds transfer without spectral distortion. □

A.4 Cumulative spectral balance with approximate steps

This subsection proves Corollary 4.3. Here $\|D_k^\top D_k - I_r\|_2 \le \varepsilon_k$ and $\delta_K = \sum_{k=1}^K \eta_k^2 \varepsilon_k$. The definitions of C_K, a_K, E_K are unchanged and use the actual steps. We assume $a_K > 0$ and $\|E_K\|_2 \le \rho a_K$, $0 \le \rho < 1$; the condition-number and rank bounds additionally require $\delta_K < a_K(1-\rho)$.

Proof. Write $B = B_K$, $C = C_K$, $a = a_K$, $E = E_K$, $\delta = \delta_K$, and, when the strict positivity condition holds, $\tau = \delta/a$.

Step 1: Isolate the accumulated orthogonalization error. Define symmetric matrices

$$F_k = D_k^\top D_k - I_r, \qquad F = \sum_{k=1}^K \eta_k^2 F_k. \tag{28}$$

The assumed stepwise error bounds imply

$$\|F\|_2 \le \sum_{k=1}^K \eta_k^2 \|F_k\|_2 \le \sum_{k=1}^K \eta_k^2 \varepsilon_k = \delta. \tag{29}$$

Expanding the Gram matrix as before now gives

$$B^\top B = aI_r + E + F. \tag{30}$$

The cross-step matrices C, E and scalar a are defined using the actual, possibly nonorthogonal D_k.

Step 2: Bound the spectrum and establish full rank. For every unit vector $x \in \mathbb{R}^r$, symmetry and equation 29 imply

$$a - \rho a - \delta \le x^\top B^\top B x \le a + \rho a + \delta. \tag{31}$$

Hence every eigenvalue of $B^\top B$ satisfies

$$a(1-\rho) - \delta \le \sigma_j(B)^2 \le a(1+\rho) + \delta, \qquad j = 1, \dots, r. \tag{32}$$

This bound remains valid even if its lower endpoint is nonpositive. For the remaining claims, assume $\delta < a(1-\rho)$. Then $\rho + \tau < 1$, and all singular values are positive. Dividing the extreme bounds in equation 32 gives

$$\kappa(B) \le \sqrt{\frac{1+\rho+\tau}{1-\rho-\tau}}. \tag{33}$$

Step 3: Bound the stable rank. Because $\mathrm{tr}(E) = 0$ and $|\mathrm{tr}(F)| \le r\|F\|_2 \le r\delta$,

$$\|B\|_F^2 = ra + \mathrm{tr}(F) \ge r(a - \delta) = ra(1-\tau). \tag{34}$$

Combining this with $\|B\|_2^2 \le a(1+\rho) + \delta = a(1+\rho+\tau)$ yields

$$\mathrm{sr}(B) \ge \frac{r(1-\tau)}{1+\rho+\tau}. \tag{35}$$

Step 4: Center the Gram matrix at its actual mean. In the approximate case, a need not equal $\|B\|_F^2/r$. Define

$$f = \frac{\mathrm{tr}(F)}{r}, \qquad \mu = a + f = \frac{\|B\|_F^2}{r}, \qquad T = E + F - fI_r. \tag{36}$$

Then

$$B^\top B = \mu I_r + T, \qquad \mathrm{tr}(T) = 0, \qquad \mu \ge a - \delta = a(1-\tau) > 0. \tag{37}$$

To bound T, use the Frobenius norm. Direct expansion gives

$$\|F - fI_r\|_F^2 = \|F\|_F^2 - \frac{\big(\mathrm{tr}(F)\big)^2}{r} \le \|F\|_F^2 \le r\delta^2. \tag{38}$$

Thus, by the triangle inequality,

$$\begin{aligned}\|T\|_F &\le \|E\|_F + \|F - fI_r\|_F \\ &\le \sqrt{r}\,\rho a + \sqrt{r}\,\delta = \sqrt{r}\,a(\rho+\tau).\end{aligned} \tag{39}$$

The centering estimate is made in the Frobenius norm; no contraction of the spectral norm under trace subtraction is required.

Step 5: Bound the entropy-based effective rank. Using equation 37, symmetry and tracelessness give

$$\mathrm{tr}\big((B^\top B)^2\big) = r\mu^2 + \|T\|_F^2. \tag{40}$$

Lemma A.2 and equation 39 therefore imply

$$\begin{aligned}\mathrm{erank}(B) &\ge \frac{(r\mu)^2}{r\mu^2 + \|T\|_F^2} = \frac{r}{1 + \|T\|_F^2/(r\mu^2)} \\ &\ge \frac{r}{1 + \left(\dfrac{\rho+\tau}{1-\tau}\right)^2}.\end{aligned} \tag{41}$$

The final inequality follows from $\mu \ge a(1-\tau) > 0$ and $\|T\|_F^2 \le ra^2(\rho+\tau)^2$.

Step 6: Transfer the bounds to the weight update. Apply Lemma A.1 to B and A. Since B has full column rank under the strict positivity condition, the condition-number, stable-rank, and entropy-effective-rank bounds hold unchanged for BA, using $\kappa_+(BA)$ for its nonzero condition number. □

B Approximate Computation and Scope of the Guarantees

B.1 Interpreting historical response

RRS uses $M^t = (X_t)^\top X_t$ and $H_{t-1} = \sum_{i<t} M^i + \epsilon I$, with $\epsilon > 0$. These unnormalized statistics weight observations according to their sample and token counts. Replacing them by per-task averages changes the relative weighting and is a different design choice. Because the backbone evolves across tasks, stored historical statistics summarize features observed at their respective collection stages. They serve as a surrogate for historical response, rather than an exact recomputation of all past features under the current backbone. Positive definiteness of H_{t-1} ensures that the spectral selection problem remains well defined; it does not imply that this surrogate exactly measures historical loss or forgetting. The method retains aggregate second-moment matrices, even though it does not replay past examples or individual historical features.

B.2 A computable characterization of an approximate RRS subspace

Let $\widetilde{M} = L^{-1}ML^{-\top} \succeq 0$ and let $\widehat{U} \in \mathbb{R}^{m\times r}$ have orthonormal columns, as returned by an approximate eigensolver. Define

$$\widehat{\mathcal{S}} = \mathrm{range}(L^{-\top}\widehat{U}), \qquad T_r = \widehat{U}^\top \widetilde{M}\widehat{U}, \qquad \theta_r = \lambda_{\min}(T_r). \tag{42}$$

For $v = L^{-\top}\widehat{U}c$ with $c \ne 0$, the same change of coordinates as in the exact proof yields

$$\frac{v^\top M v}{v^\top H v} = \frac{c^\top T_r c}{c^\top c}. \tag{43}$$

Consequently, the worst-direction relative response of the computed subspace is exactly θ_r. Euclidean QR of $L^{-\top}\widehat{U}$ preserves this value by preserving the subspace. If λ_r is the r-th largest eigenvalue of

$\widetilde{M}$, max–min optimality gives $0 \le \theta_r \le \lambda_r$. This characterizes the approximate subspace without assuming exact RSVD optimality.

A sufficient approximation certificate is available from $\widehat{M} = \widehat{U} T_r \widehat{U}^\top$. If $\|\widetilde{M} - \widehat{M}\|_2 \le \zeta$, then

$$\max\{0, \lambda_r - \zeta\} \le \theta_r \le \lambda_r. \tag{44}$$

Indeed, $\widehat{M}$ is positive semidefinite, its r-th largest eigenvalue is θ_r (including zeros), and Weyl's eigenvalue inequality bounds its difference from λ_r by ζ. This certificate is sufficient and may be conservative; no numerical value of ζ is assumed here. A small invariant-subspace residual alone does not establish that the computed subspace is the leading one.

B.3 Interpretation and scope of the cumulative guarantees

Why a cross-step condition is necessary. Individually orthogonal steps can cancel along selected directions. For example, take $n = r = 2$, $K = 2$, $\eta_1 = \eta_2 = 1$, and

$$D_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{45}$$

Both satisfy $D_k^\top D_k = I_2$, but

$$B_2 = -(D_1 + D_2) = \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix} \tag{46}$$

is rank deficient. Here $a_2 = 2$ and $E_2 = \mathrm{diag}(2, -2)$, so $\|E_2\|_2 / a_2 = 1$. The strict anisotropy assumption correctly excludes this example. The theorem therefore provides a conditional spectral guarantee, not an unconditional consequence of stepwise orthogonalization.

Exact and approximate orthogonalization. If $\tau_K = 0$, then $\delta_K = 0$ and the accumulated error matrix F vanishes. The corollary then recovers the theorem's bounds in terms of ρ; retaining the exact Frobenius norm of E_K also recovers its sharper effective-rank bound. A finite polynomial orthogonalization does not by itself guarantee $\rho + \tau_K < 1$. Similarly, a rank-deficient step necessarily has $\|D_k^\top D_k - I_r\|_2 \ge 1$. The strict cumulative condition must be checked or justified for the actual sequence of updates.

Within-task versus across-task accumulation. The analysis fixes a single input basis A and studies $B_K A = -\sum_{k=1}^K \eta_k D_k A$ within that task. It does not directly apply to the across-task sum $\sum_{t=1}^T B_t A_t$ when the bases A_t vary. The results characterize update geometry; they do not independently establish a bound on task loss, classification error, or forgetting.

B.4 Checking the cumulative conditions without pairwise step storage

The cumulative conditions can be evaluated from the actual factor steps. Set $D_k = -\Delta B_k / \eta_k$ for an additive update $\Delta B_k = B_k - B_{k-1}$ with $B_0 = 0$ and $\eta_k > 0$. All step scaling must be included consistently; if the implementation contains additional terms, they cannot silently be omitted. Accumulate

$$S_K = \sum_{k=1}^K \eta_k^2 D_k^\top D_k, \quad s_K = \sum_{k=1}^K \eta_k^2, \quad \delta_K = \sum_{k=1}^K \eta_k^2 \|D_k^\top D_k - I_r\|_2. \tag{47}$$

The identity

$$C_K = B_K^\top B_K - S_K \tag{48}$$

recovers exactly the cross-step matrix in the theorem without retaining all pairs of steps. Compute $a_K = s_K + \mathrm{tr}(C_K)/r$ and $E_K = C_K - \mathrm{tr}(C_K) I_r / r$. For $a_K > 0$, define

$$\widehat{\rho}_K = \frac{\|E_K\|_2}{a_K}, \qquad \widehat{\tau}_K = \frac{\delta_K}{a_K}. \tag{49}$$

The strict certificate holds whenever $\widehat{\rho}_K + \widehat{\tau}_K < 1$; the corollary can then be evaluated using these quantities. Cases with $a_K \le 0$ or a failed strict inequality must be counted as uncertified rather than discarded. Failure of this sufficient certificate does not imply poor accuracy or rank deficiency. These expressions specify a diagnostic procedure; the accuracy comparisons below do not themselves establish that the certificate holds along training trajectories.

Table 4: Comparison of subspace construction and spectral control on ImageNet-A with 20 incremental sessions. A_{Last} and A_{Avg} are reported in %.

Method	$A_{\text{Last}} \uparrow$	$A_{\text{Avg}} \uparrow$
Trainable A, B + Muon	51.34 ± 0.35	62.46 ± 0.33
Fixed random orthonormal A + Muon	54.32 ± 0.64	64.53 ± 0.51
Fixed current-task PCA A + Muon	52.45 ± 0.34	63.56 ± 0.67
Fixed historical-isolation A + Muon	54.67 ± 0.25	64.76 ± 0.47
EBLoRA	52.52 ± 1.01	62.16 ± 0.82
MuLoRA (RRS + SSB)	$\mathbf{55.86} \pm 0.61$	$\mathbf{65.61} \pm 0.64$

C Mechanism Comparisons on 20S-ImageNet-A

C.1 Definitions of the compared variants

The comparison separates choices of input subspace from the location at which spectral control is applied. A fixed basis is held constant within each task and reconstructed for the next task.

Trainable A, B + Muon. Both LoRA factors are optimized using Muon. The induced weight step is $\Delta BA + B\Delta A + \Delta B\Delta A$, so controlling the spectra of the two factor steps separately does not ensure spectral control of their combined weight step. This variant also has more trainable adapter parameters: $r(m+n)$ rather than nr per adapted matrix.

Fixed random orthonormal A + Muon. A random row-orthonormal basis satisfies $AA^\top = I_r$ and only B is trained. This retains exact factor-to-weight spectral transfer without data-dependent subspace selection.

Fixed current-task PCA A + Muon. Let U_r contain the leading r eigenvectors of M^t and set $A = U_r^\top$. This selects the strongest current-task second-moment directions without historical information. Here "PCA" denotes uncentered second-moment spectral selection, consistent with M^t; centering the features would define a different variant.

Fixed historical-isolation A + Muon. Let U_H be an orthonormal basis for the retained historical feature subspace, and let Q_H span its orthogonal complement. If V_r contains the leading r eigenvectors of $Q_H^\top M^t Q_H$, set $A = (Q_H V_r)^\top$. This maximizes current-task response subject to hard historical isolation and requires $\dim(\text{range}(Q_H)) \geq r$. The retained historical dimension determines the feasible space.

EBLoRA. EBLoRA constrains the adapter's nonzero singular values through orthonormal factors and a shared scalar magnitude. The distinction is the object of spectral control: an adapter-spectrum constraint versus MuLoRA's control of individual momentum steps.

MuLoRA (RRS + SSB). RRS selects directions by current-task response relative to historical response, followed by Euclidean orthonormalization. SSB updates only B within this fixed basis. It does not require the accumulated adapter to have equal singular values, and thus retains spectral freedom across optimization steps.

C.2 Results and interpretation

Table 4 reports the mechanism comparison. MuLoRA attains the highest mean final and average accuracy. Relative to the fixed random orthonormal basis, RRS improves A_{Last} and A_{Avg} by 1.54 and 1.08 percentage points, respectively. Its gains over current-task PCA are 3.41 and 2.05 points, suggesting that strong current-task feature response alone is insufficient to select an effective continual adaptation space in this setting. Relative to historical isolation, the gains are 1.19 and 0.85 points, supporting graded relative-response selection over strictly excluding retained historical directions.

MuLoRA exceeds jointly trainable factors with Muon by 4.52 and 3.15 points and EBLoRA by 3.34 and 3.45 points. These comparisons support the effectiveness of the complete design. They do not establish that fixed orthonormal parameterization is necessary, or isolate a causal effect of spectral balancing: trainable-factor counts, optimization dynamics, and subspace choice can differ across variants. Together with the main-text RRS/SSB component ablation, the results support combining history-aware subspace selection and stepwise spectral balancing. The reported mean differences are not claims of statistical significance; paired runs are needed to quantify their uncertainty. Overall and average accuracy alone do not identify whether a gain arises from improved new-task learning or reduced forgetting.

D Reproducibility Details and Limitations

D.1 Reported experimental configuration

The main experimental configuration uses a ViT-B/16 backbone pretrained on ImageNet-21K, adapter rank $r = 32$, a cosine classifier, and a single NVIDIA RTX 4090 GPU. Each task is trained for 20 epochs with batch size 48 and a cosine learning-rate schedule. The reported initial learning rate is 10^{-4} on ImageNet-R and 2×10^{-4} on the other benchmarks, including ImageNet-A. The main comparison reports averages over three random seeds.

D.2 Memory and computational accounting

For an adapted matrix of size $n \times m$, the trainable factor $B \in \mathbb{R}^{n \times r}$ has nr parameters, while the fixed basis requires mr stored entries during a task. A dense historical second-moment matrix requires $O(m^2)$ persistent storage per distinct tracked input representation. Thus absence of replay does not imply absence of historical state. Statistics may be shared only where the corresponding input features are identical.

For N_t collected feature vectors, directly forming M^t costs $O(N_t m^2)$ operations and can be accumulated in batches without retaining all feature vectors. A dense Cholesky factorization costs $O(m^3)$; triangular solves, approximate eigenspace extraction, and QR contribute additional task-initialization costs. For $B \in \mathbb{R}^{n \times r}$, each polynomial orthogonalization iteration based on $Y^\top Y$ costs $O(nr^2 + r^3)$, with $r \leq n$. These operation counts do not establish a wall-clock advantage. End-to-end efficiency depends on feature collection and subspace construction as well as optimizer steps.

D.3 Scope of empirical and theoretical conclusions

The spectral-transfer identity is exact for a row-orthonormal basis, whereas the implemented polynomial orthogonalization and eigenspace selection are approximate. Cumulative guarantees depend on measured or otherwise justified cross-step conditions, and apply within a task with a fixed basis. They are not bounds on cross-task forgetting, classification error, or generalization.

The mechanism comparison covers one 20-session benchmark setting. Broader main-text accuracy results do not replace matched mechanism comparisons across additional settings. A norm-matched momentum control would help distinguish update-scale effects from spectral reshaping. Separate old-class and new-class measurements, with the prediction label space specified, would further distinguish retention from adaptation. Spectral plots should identify the update matrix, layer aggregation, normalization, and numerical rank threshold; κ_+ should not be interpreted without rank-sensitive statistics.