WAS-LoRA: Output-Sensitive Subspace Selection for Post-Training LoRA Compression

Anonymous authors
Paper under double-blind review

Abstract

We study post-training LoRA compression to reuse trained adapters under tighter rank budgets at deployment, without requiring an additional fine-tuning stage. Small reconstruction errors in adapter weights or the local LoRA branch outputs, however, do not necessarily ensure that the compressed model's final logits remain close to those of the original adapted model. We propose WAS-LoRA (Whitened Active Subspace Low-Rank Adaptation), which guides active-subspace selection using the sensitivity of a selected teacher output downstream of the adapter's bottleneck. It constructs a Jacobian Gram matrix that measures the output sensitivity in the adapter's whitened bottleneck coordinates. The leading eigenvectors of this Jacobian Gram matrix span the retained subspace, from which we directly construct standard lower-rank LoRA factors. We prove that, for a fixed retained rank, this subspace minimizes a Jacobian-based measure of output sensitivity along discarded bottleneck directions. Experiments on GLUE classification tasks across five different rank budgets demonstrate that WAS-LoRA reduces geometric-mean centered relative logit error by 30.1% and 15.8% compared with the activation-aware and fixed-basis baselines, respectively. Task-accuracy differences among methods are concentrated at the tightest rank budgets.

1 Introduction

Low-Rank Adaptation (LoRA) has become a standard approach for adapting large pretrained models by keeping their original weights frozen and learning a low-rank update instead (Hu et al., 2022). A LoRA module parameterizes the update as $\Delta W = \gamma BA$, where $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ are trainable factors, r is the adapter rank, and γ is a scaling factor. This factorization introduces an r-dimensional bottleneck in the adapter branch, reducing the number of trainable parameters and per-task storage. Yet the rank a deployment can afford can be constrained by the serving system, as many adapters share a backbone and compete for limited cache capacity, while each adapter's memory footprint grows linearly with r (Sheng et al., 2024; Wu et al., 2024). Retraining a lower-rank adapter is not always available: the original adaptation data or training procedure may be inaccessible, and the deployment rank budget may only be known after adaptation. Training-time multi-rank methods such as DyLoRA support a range of ranks from a single adaptation run (Valipour et al., 2023), addressing deployment flexibility when the adaptation process can be controlled, whereas our setting starts from an already trained adapter. We therefore study how to reduce a trained adapter's rank while preserving the behavior of the original adapted model (the teacher). The teacher remains fixed, and compression uses unlabeled calibration inputs without recovery fine-tuning of the compressed model. Then the question is not how to learn a new adaptation, but what to retain from an existing one.

We organize the related literature around two complementary questions: what compression aims to preserve and how the retained subspace is selected. Spectral post-hoc methods retain directions that best reconstruct the LoRA update in weight space, using singular structure for rank truncation or allocation (Vasani et al., 2025). Activation-aware low-rank approximation additionally incorporates representative inputs, as in the local whitening-based reconstruction objective used by dense-model methods such as SVD-LLM (Wang et al., 2025b). Yet local errors of similar magnitude can be amplified or suppressed differently by later layers. Output-aware dense-model methods such as IO-SVD incorporate predictive sensitivity through input-output whitening (Abbasi et al., 2026); Figure 1 contrasts these three preservation targets. For subspace selection, AdaLoRA, IGU-LoRA, and GoRA use gradient-derived signals to score components or allocate rank during adaptation or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

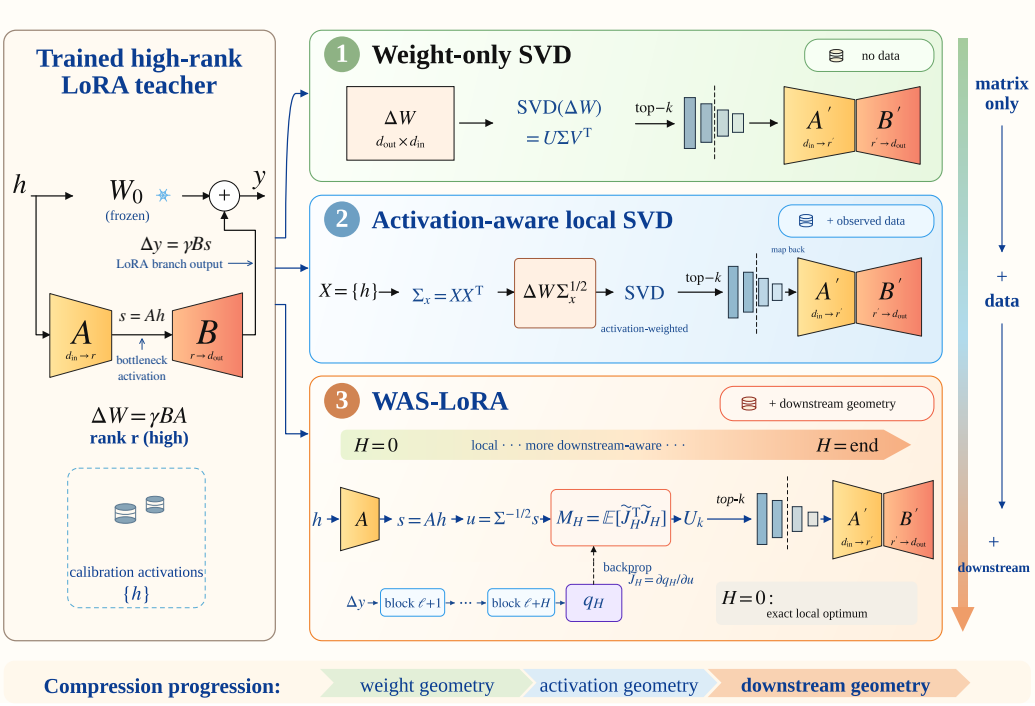


Figure 1: **Preservation targets for post-training LoRA compression.** Weight-space SVD reconstructs the update ΔW (top), activation-aware SVD reconstructs the local LoRA branch output $\Delta y = \gamma Bs$ on calibration inputs with $s = Ah$ (middle), and WAS-LoRA selects directions in the whitened bottleneck activation $u = \Sigma^{-1/2}s$ using the sensitivity of a downstream endpoint q_H (bottom). Here $H = 0$ denotes the local branch output, integer $H \geq 1$ denotes relative propagation depth to deeper hidden states, and $H = \text{end}$ denotes the final logits. The selected directions are folded back into standard LoRA factors, requiring no custom inference operator.

initialization (Zhang et al., 2023; Cui et al., 2026; He et al., 2025), while post-training refinement methods such as Spectral Surgery use calibration-loss gradients for post-training singular-value reweighting while retaining the adapter's singular directions (Tian et al., 2026). For a frozen adapter, however, rescoring a basis chosen for local reconstruction need not recover the subspace obtained when downstream sensitivity is allowed to select new linear combinations of its vectors.

We introduce WAS-LoRA (Whitened Active Subspace Low-Rank Adaptation), a post-training compression framework that selects a retained subspace in a trained adapter's bottleneck activation space using the sensitivity of a chosen teacher output. The preservation target can be a local LoRA branch output, a hidden state, or the final logits. WAS-LoRA whitens the adapter's bottleneck activations using unlabeled calibration inputs and constructs a Jacobian Gram matrix that measures sensitivity of the selected teacher output to perturbations in this whitened activation space. The leading eigenvectors of this Jacobian Gram matrix span the retained subspace, from which we directly construct standard lower-rank LoRA factors. No recovery training or custom inference operators are required. We prove that, for a fixed retained rank, this subspace minimizes a Jacobian-based measure of output sensitivity along discarded directions in this space. Across five GLUE tasks, with eight rank-16 teachers per task and five retained-rank budgets, WAS-LoRA consistently improves teacher-function fidelity across the evaluated rank budgets. Its geometric-mean centered relative logit error is 30.1% lower than that of the local activation-aware control baseline and 15.8% lower than that of the information-matched fixed-basis control baseline. Teacher-function fidelity and task accuracy need not move in lockstep: at moderate budgets, methods can remain close in accuracy even when their output fidelity differs substantially, whereas task-accuracy differences become most visible under tight rank constraints. We therefore treat teacher-function fidelity as the primary quantity in this label-free, zero-recovery compression setting, with task accuracy as a secondary downstream measure.

Our main contributions are:

- **Flexible output-preservation formulation.** We formulate post-training LoRA compression using a Jacobian-based sensitivity criterion for preserving a chosen teacher output. The preservation target can range from local LoRA branch outputs to final logits.
- **Tractable subspace construction with theoretical guarantees.** We develop WAS-LoRA, an active-subspace-inspired construction that selects a retained subspace in a trained adapter's whitened bottleneck activation space using the leading eigenvectors of a Jacobian Gram matrix with theoretical guarantees.
- **Fixed-teacher evaluation.** Across five retained-rank budgets, experiments on GLUE tasks show consistently improved teacher-function fidelity, with task-accuracy differences concentrated at the tightest rank budgets. Comparisons with fixed-basis controls demonstrate gains beyond downstream rescoring. We also quantify one-shot calibration costs and evaluate downstream-sensitivity scoring through held-out deletion tests on fixed SVD candidates.

2 Motivation

2.1 When Local Geometry Is the Wrong Objective

A locally dominant adapter direction need not affect the model's final output. Consider the rank-two teacher $f_{\widetilde{W}}(x) = \mathbf{1}^\top \mathrm{ReLU}((W_0 + \widetilde{W})x + b)$, where $W_0 = I_2$, $b = (-20, 3)^\top$, $\widetilde{W} = \mathrm{diag}(10, 1)$, and $x \sim \mathrm{Unif}([-1, 1]^2)$.

As Figure 2 shows, the upper ReLU is always inactive while the lower one is always active, giving $f_{\widetilde{W}}(x) = 3 + 2x_2$.

Rank-one SVD retains $\widetilde{W}_{\mathrm{SVD}} = \mathrm{diag}(10, 0)$. Because $\mathbb{E}[xx^\top] = I_2/3$, this also minimizes the activation-aware local reconstruction error $\mathbb{E}\|(\widetilde{W} - U)x\|_2^2$ over rank-one U. Yet $f_{\widetilde{W}_{\mathrm{SVD}}}(x) = 3 + x_2$. Instead, $\widetilde{W}_{\mathrm{alt}} = \mathrm{diag}(0, 1)$ retains the smaller direction and reproduces the teacher output exactly. Thus, for $\mathcal{L}(U) = \mathbb{E}|f_{\widetilde{W}}(x) - f_U(x)|^2$, $\mathcal{L}(\widetilde{W}_{\mathrm{SVD}}) = 1/3$ while $\mathcal{L}(\widetilde{W}_{\mathrm{alt}}) = 0$. The locally optimal direction can therefore be suppressed by downstream computation, motivating compression based on the teacher's output. Further counterexamples, including higher-dimensional rank compression and a softmax-attention realization, appear in Section B.

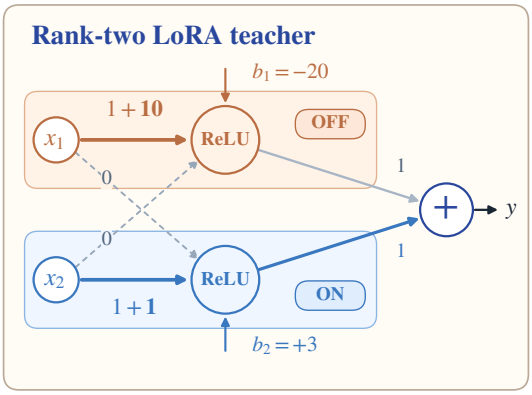


Figure 2: **A locally dominant adapter direction can be silent.** Edge labels give total weights (backbone + adapter); the two dashed crossed edges have weight zero.

2.2 From Output Preservation to a Sensitivity Criterion

The preceding example motivates preserving the teacher's final logits, rather than only adapter weights or local LoRA branch outputs. For a trained LoRA module indexed by ℓ with input h_ℓ, the update is $\Delta y_\ell = \gamma B_\ell A_\ell h_\ell$, where $s_\ell(x) = A_\ell h_\ell(x) \in \mathbb{R}^{r_\ell}$ is the bottleneck activation. Compression reduces this bottleneck's width. For the fixed module ℓ at block index $b_\ell \in \{0, \ldots, L-1\}$, we use H to index the propagation endpoint: $H = 0$ denotes the local LoRA branch output, integer $H \geq 1$ denotes the output of block $\min(b_\ell + H, L-1)$, and $H = \mathrm{end}$ denotes the final logits. Let $f_{\ell \to H}(x, s)$ denote the corresponding endpoint obtained by replacing module ℓ's bottleneck activation with $s \in \mathbb{R}^{r_\ell}$. At the teacher activation $s_\ell(x)$, $q_H(x) = f_{\ell \to H}(x, s_\ell(x))$. We define the corresponding Jacobian with respect to the bottleneck activation as

$$J_\ell^{(H)}(x) = \left. \frac{\partial f_{\ell \to H}(x, s)}{\partial s} \right|_{s = s_\ell(x)}.$$

This Jacobian captures the first-order sensitivity of the selected output to perturbations in module ℓ's bottleneck activation.

Let T be a linear compression map with $\mathrm{rank}(T) \leq k$. Replacing $s_\ell(x)$ with $Ts_\ell(x)$ while leaving all other modules unchanged gives the single-module output-preservation loss

$$\mathcal{L}_H(T) = \mathbb{E}_x \left\| f_{\ell \to H}(x, s_\ell(x)) - f_{\ell \to H}(x, Ts_\ell(x)) \right\|_2^2, \tag{1}$$

where the expectation is over the input distribution. This loss directly measures the change in the selected teacher output caused by compression. In general, optimizing it requires repeated downstream evaluations for candidate maps, while downstream nonlinearities complicate the objective. When multiple modules are compressed, their effects can also interact. Rather than directly optimizing this finite nonlinear loss, we seek a tractable criterion for choosing which bottleneck directions should be retained. Active Subspaces provides such a sensitivity-based principle.

An active subspace motivation The classical Active Subspaces (AS) framework motivates sensitivity-based subspace selection through low-dimensional function approximation (Constantine et al., 2014; Zahm et al., 2020). Let $Z \in \mathbb{R}^d$ be a random input and $V \in \mathbb{R}^{d \times k}$ have orthonormal columns. For a differentiable function f with square-integrable values and derivatives, the L^2-optimal approximation of $f(Z)$ based only on the retained coordinates $V^\top Z$ is the conditional expectation $g_V^\star(V^\top Z) = \mathbb{E}_Z[f(Z) \mid V^\top Z]$. Assume that the conditional distributions along the discarded directions, given $V^\top Z$, satisfy a Poincaré inequality with a constant C_P uniform over the retained coordinates and the subspaces considered. Applying this inequality to each component of f at fixed $V^\top Z$, and then averaging over the retained coordinates, yields the subspace Poincaré bound

$$\mathbb{E}_Z \| f(Z) - g_V^\star(V^\top Z) \|_2^2 \leq C_P \, \mathbb{E}_Z \| J_f(Z)(I - VV^\top) \|_F^2, \tag{2}$$

where $J_f(Z)$ is the input Jacobian of f evaluated at Z, and $I - VV^\top$ is the orthogonal projector onto the discarded subspace. The Jacobian term measures the average squared output sensitivity along discarded directions, motivating subspace selection by minimizing the right-hand side of Eq. (2). This sensitivity criterion guides our direction selection, but the bound concerns the optimal conditional-expectation predictor based on $V^\top Z$. Our compressed model instead projects the realized bottleneck activation and reuses the original downstream network; its output error is not directly controlled by this bound.

3 WAS-LoRA Algorithm

Given a frozen LoRA teacher and unlabeled calibration inputs, WAS-LoRA constructs lower-rank adapters under a total rank budget. The details on sampling the calibration inputs are described in the next section. The calibration set $\mathcal{D}$ contains a collection of inputs from the adapted task. Our primary configuration targets the teacher's final logits. Algorithm 1 summarizes the procedure: collect teacher statistics, construct whitened sensitivity matrices, allocate retained directions, and form the compressed LoRA factors. For each module, $\widehat{\Sigma}_\ell$ denotes the empirical bottleneck second moment and $\widehat{C}_{H,\ell}$ the raw Jacobian Gram matrix of the output q_H. Their whitened product $\widehat{M}_{H,\ell}$ determines the retained eigenvectors.

Calibration and sensitivity construction. Algorithm 1 (lines 1–6) estimates bottleneck second-moment matrices $\widehat{\Sigma}_\ell$ and Jacobian Gram matrices $\widehat{C}_{H,\ell}$ in the original bottleneck coordinates, then combines them to construct sensitivity matrices $\widehat{M}_{H,\ell}$ in whitened coordinates. We first present the unregularized population construction, with $\mathbb{E}_x$ denoting expectation under the calibration input distribution. Omitting the module index, we define

$$\Sigma = \mathbb{E}_x[s(x)s(x)^\top], \qquad u = \Sigma^{-1/2} s, \qquad \widetilde{J}_H(x) = J_H(x)\Sigma^{1/2}. \tag{3}$$

Here $J_H(x)$ is the Jacobian of the selected teacher output q_H with respect to the original bottleneck coordinates $s = Ah$, and Σ is an uncentered second-moment matrix. The matrix $\widetilde{J}_H(x)$ is the Jacobian with respect to the whitened bottleneck coordinates u, obtained by applying the chain rule to $s = \Sigma^{1/2}u$. If Σ is singular, the unregularized construction is restricted to $\mathrm{range}(\Sigma)$. Inverse maps, identity matrices, and projectors are then understood on this subspace, where $\mathbb{E}[uu^\top] = I$. A rank-k

Algorithm 1 WAS-LoRA compression

Require: Fixed teacher with LoRA weights $\{A_\ell, B_\ell, \gamma_\ell\}$, unlabeled calibration data $\mathcal{D}$, selected teacher output q_H, feasible total rank budget K, minimum per-module ranks m_ℓ (primary WAS: $m_\ell = 1$).

1: Process batches from $\mathcal{D}$ with the fixed teacher and record bottleneck activations $s_\ell = A_\ell h_\ell$ at non-padding positions.
2: Accumulate activation statistics $\widehat{\Sigma}_\ell$ from the recorded bottleneck activations.
3: Estimate Jacobian Gram matrices $\widehat{C}_{H,\ell}$ in the original bottleneck coordinates using gradients of individual coordinates of q_H or stochastic-probe VJPs (E.3).
4: **for** each module ℓ **do**
5: Compute $S_\ell = \widehat{\Sigma}_{\ell,\lambda}^{1/2}$ and S_ℓ^{-1} as specified in Eq. (73).
6: Form the Jacobian Gram matrix in whitened bottleneck coordinates: $\widehat{M}_{H,\ell} = S_\ell \widehat{C}_{H,\ell} S_\ell$.
7: Compute the eigendecomposition $\widehat{M}_{H,\ell} = U_\ell \Lambda_\ell U_\ell^\top$ with eigenvalues sorted in descending order. Initialize the retained rank of module ℓ as $k_\ell \leftarrow m_\ell$.
8: **end for**
9: Pool the eigenvalues beyond the first m_ℓ in each module. Select the $K - \sum_\ell m_\ell$ largest eigenvalues. For each selected eigenvalue from module ℓ, update $k_\ell \leftarrow k_\ell + 1$.
10: For each module ℓ, construct $(A_{\ell,k_\ell}, B_{\ell,k_\ell})$ from the leading k_ℓ columns of U_ℓ using Eq. (7), with the same whitening maps and unchanged γ_ℓ.
11: **return** standard lower-rank LoRA weights $\{A_{\ell,k_\ell}, B_{\ell,k_\ell}, \gamma_\ell\}$.

orthogonal projector P in these whitened coordinates corresponds to the original-coordinate map $T_P = \Sigma^{1/2} P \Sigma^{-1/2}$.

Let $C_H = \mathbb{E}_x[J_H(x)^\top J_H(x)]$ be the corresponding Jacobian Gram matrix in the original bottleneck coordinates. The Jacobian Gram matrix in whitened coordinates is

$$M_H = \mathbb{E}_x[\widetilde{J}_H(x)^\top \widetilde{J}_H(x)] = \Sigma^{1/2} C_H \Sigma^{1/2}. \tag{4}$$

In practice, $\widehat{\Sigma}_\ell$ is accumulated from bottleneck activations at non-padding positions in the calibration set $\mathcal{D}$. The corresponding $\widehat{C}_{H,\ell}$ is estimated using gradients of individual coordinates of q_H or stochastic-probe VJPs. For numerical stability, the implementation uses the ridge-regularized second-moment $\widehat{\Sigma}_{\ell,\lambda}$ and forms $\widehat{M}_{H,\ell} = S_\ell \widehat{C}_{H,\ell} S_\ell$, where $S_\ell = \widehat{\Sigma}_{\ell,\lambda}^{1/2}$.

Spectral selection and rank allocation. Algorithm 1 (lines 7–9) eigendecomposes each $\widehat{M}_{H,\ell}$, with eigenvalues $\lambda_{\ell,j}$ sorted in descending order. It first reserves each module's leading m_ℓ directions, setting $k_\ell = m_\ell$, then selects the $K - \sum_\ell m_\ell$ largest remaining eigenvalues across modules. Each selection increases the corresponding k_ℓ by one. Module ℓ retains its leading k_ℓ eigenvectors U_{ℓ,k_ℓ}, implicitly defining the projector $P_\ell = U_{\ell,k_\ell} U_{\ell,k_\ell}^\top$. This construction retains a leading prefix of directions within each module and allocates the total rank budget across modules.

To justify this procedure, consider a rank-k orthogonal projector P onto a retained subspace in whitened coordinates. Following the classical AS motivation in Eq. (2), we adopt its discarded-direction Jacobian energy as our tractable subspace-selection criterion. It measures sensitivity along discarded coordinate directions. Define the sensitivity along the discarded directions as

$$R_H^{\mathrm{w}}(P) = \mathbb{E}_x \|\widetilde{J}_H(x)(I - P)\|_F^2 = \mathrm{tr}((I - P) M_H). \tag{5}$$

As $\mathrm{tr}(M_H)$ is fixed, minimizing this sensitivity surrogate is equivalent to maximizing $\mathrm{tr}(P M_H)$.

Theorem 1 (Optimal discarded sensitivity)**.** *Working on* $\mathrm{range}(\Sigma)$*, let* $M_H = U \Lambda U^\top$ *be the eigendecomposition with eigenvalues in descending order. For* $0 \le k \le \mathrm{rank}(\Sigma)$*, let* U_k *contain the leading* k *eigenvectors. Then*

$$U_k U_k^\top \in \arg \min_{\substack{P = P^\top = P^2 \\ \mathrm{rank}(P) = k}} R_H^{\mathrm{w}}(P), \qquad \min_{\substack{P = P^\top = P^2 \\ \mathrm{rank}(P) = k}} R_H^{\mathrm{w}}(P) = \sum_{i > k} \lambda_i(M_H). \tag{6}$$

For each fixed empirical matrix $\widehat{M}_{H,\ell}$, the theorem implies that increasing k_ℓ by one reduces the minimum discarded sensitivity by $\lambda_{\ell,k_\ell+1}$. With one rank unit per retained direction, the allocation

above therefore exactly minimizes the module-wise additive surrogate $\sum_\ell \mathrm{tr}((I - P_\ell)\widehat{M}_{H,\ell})$ subject to $\sum_\ell k_\ell = K$ and $k_\ell \geq m_\ell$.

Constructing compressed adapters. After rank allocation, Algorithm 1 (lines 10) converts the retained eigenvectors into standard low-rank LoRA factors. Let U_{ℓ,k_ℓ} contain the leading k_ℓ eigenvectors of $\widehat{M}_{H,\ell}$ and set $P_\ell = U_{\ell,k_\ell}U_{\ell,k_\ell}^\top$. Using the same maps S_ℓ as in the sensitivity construction, we obtain

$$A_{\ell,k_\ell} = U_{\ell,k_\ell}^\top S_\ell^{-1} A_\ell, \qquad B_{\ell,k_\ell} = B_\ell S_\ell U_{\ell,k_\ell}. \tag{7}$$

The resulting branch satisfies $\gamma_\ell B_{\ell,k_\ell}A_{\ell,k_\ell}h_\ell = \gamma_\ell B_\ell S_\ell P_\ell S_\ell^{-1}A_\ell h_\ell$. The projection is therefore absorbed into the standard LoRA factors, without introducing an additional inference operator. We preserve the original scaling factor γ_ℓ. For $k_\ell > 0$, if the implementation expresses the compressed adapter's scaling as $\alpha_{\ell,k_\ell}/k_\ell$, we set $\alpha_{\ell,k_\ell} = \gamma_\ell k_\ell$. All output-gradient computations are confined to calibration, and the resulting adapters are deployed without recovery training.

We next establish that, with the local LoRA branch output selected ($H = 0$), the ridge-free spectral selection and factor construction solve the activation-aware rank-k reconstruction problem exactly.

Theorem 2 (Exact local activation-aware optimality)**.** *Let $X = n^{-1/2}[h_1, \ldots, h_n]$ collect calibration input activations for a fixed LoRA module. For $0 \leq k \leq \mathrm{rank}(AX)$, the ridge-free construction with $H = 0$ produces an update $W_k = \gamma B_k A_k$ of rank at most k satisfying*

$$W_k X = (\gamma BAX)_k, \qquad W_k \in \arg\min_{\mathrm{rank}(\widetilde{W})\leq k} \|(\gamma BA - \widetilde{W})X\|_F^2. \tag{8}$$

Here, $(Y)_k$ denotes a best rank-k approximation of a matrix Y under the Frobenius norm, obtained by truncated SVD.

We denote this LoRA-specific activation-aware baseline by $H = 0$. It shares the whitening principle of SVD-LLM, but not its full dense-weight compression pipeline (Wang et al., 2025b). Selecting a downstream teacher output therefore changes the preservation criterion, rather than improving the solution of the same local objective (Figure 1). The direction reversal in Sections 2.1 and B.4 illustrates this distinction. The compact-SVD construction and proof are given in Section A.4.

LoRA factors are not unique: for any invertible R, replacing (A, B) with (RA, BR^{-1}) leaves the teacher update γBA unchanged. The following proposition shows that, in the ridge-free setting, WAS-LoRA also yields the same compressed branch outputs from these equivalent factorizations.

Proposition 3 (Ridge-free LoRA gauge invariance)**.** *For any invertible refactorization $A \mapsto RA$, $B \mapsto BR^{-1}$, ridge-free* WAS-LoRA *preserves the supported eigenvalues of M_H. For a fixed retained rank and matched selections within tied eigenspaces, the compressed branch outputs are unchanged whenever $Ah \in \mathrm{range}(\Sigma)$. If $\Sigma \succ 0$, the full compressed update $\gamma B_k A_k$ is invariant.*

The proof and practical ridge convention are given in Sections A.5 and E.3.

4 Experiments

In this section, we conduct experiments to evaluate WAS-LoRA across retained-rank budgets, measuring teacher-function fidelity and task accuracy.

Matched fixed-teacher evaluation. We evaluate five GLUE tasks (SST-2, QNLI, MRPC, RTE, and MNLI) (Wang et al., 2018), with eight independently trained rank-16 RoBERTa-base teachers per task (Liu et al., 2019). Each teacher uses query/value LoRA in all 12 blocks, yielding 24 modules. Average retained ranks $\bar{r} \in \{2, 3, 4, 6, 8\}$ correspond to total rank budgets $K = 24\bar{r}$. All methods compress the same fixed teacher, preserving its task head and original LoRA scaling, without recovery training. For each task, we reserve a fixed calibration set from its training split, disjoint from teacher training, checkpoint selection, and evaluation. The calibration sizes are 2,048 for SST-2, QNLI, and MNLI, 512 for MRPC, and 384 for RTE. Calibration-based methods share these inputs across rank budgets. Partition, construction, and seed details are provided in Sections C.3 and E.

Baselines. We compare against Uniform-SVD, PARA (Kumaravelu et al., 2026), and Post-Squeeze (Vulić et al., 2026), and the exact ridge-free local baseline Act-SVD (Theorem 2). Output-scored

Act-SVD uses the same raw-logit sensitivity as WAS-LoRA but restricts selection to fixed Act-SVD directions. AdaLoRA-score and IGU-LoRA-score are supervised frozen-teacher scoring controls, not reproductions of the original training procedures (Zhang et al., 2023; Cui et al., 2026). Construction and allocation details are in Sections C.2 and E.4.

Evaluation metrics. For teacher logits $Z \in \mathbb{R}^{n \times C}$ and compressed logits $\widehat{Z}$, let $H_C = I_C - \mathbf{1}\mathbf{1}^\top / C$. Our primary fidelity metric is centered relative logit error,

$$e_{\mathrm{ctr}} = \frac{\|(\widehat{Z} - Z)H_C\|_F}{\|ZH_C\|_F + 10^{-12}}.$$

Centering removes class-independent additive logit shifts, which leave the softmax distribution unchanged. We additionally report teacher-to-compressed KL divergence as a complementary probability-space fidelity measure, together with task accuracy.

4.1 Does WAS preserve the trained teacher function better?

Table 1 reports geometric-mean ratios of WAS-LoRA error to baseline error, with values below 1 favoring WAS-LoRA. Each task uses the task-specific calibration size described above. Task-level entries aggregate five rank budgets and eight paired teachers, while overall entries weight the five tasks equally. All task-level centered relative logit error ratios favor WAS-LoRA. The five-task aggregate also favors WAS-LoRA when evaluated separately at each rank budget.

Table 1: Geometric-mean ratios of WAS-LoRA error to baseline error. Values below 1 favor WAS-LoRA. Overall logit-error and KL-divergence ratios additionally weight the five tasks equally.

	Centered relative logit error						KL
Comparator	SST-2	QNLI	MRPC	RTE	MNLI	Overall	Overall
Uniform-SVD	0.687	0.558	0.443	0.378	0.608	0.523	0.259
PARA	0.627	0.613	0.346	0.310	0.701	0.492	0.270
Post-Squeeze	0.687	0.558	0.443	0.378	0.608	0.523	0.259
Act-SVD	0.786	0.789	0.605	0.526	0.847	0.699	0.455
Output-scored Act-SVD	0.883	0.885	0.732	0.802	0.924	0.842	0.716
AdaLoRA-score	0.791	0.815	0.733	0.724	0.864	0.784	0.644
IGU-LoRA-score	0.731	0.756	0.746	0.692	0.836	0.751	0.632

Overall, WAS-LoRA reduces geometric-mean centered relative logit error by **30.1%** relative to the exact local Act-SVD baseline and by **15.8%** relative to Output-scored Act-SVD. **All overall KL-divergence ratios are also below one**. These comparisons support the value of using downstream sensitivity to determine new directions, beyond optimal local reconstruction and fixed-basis rescoring. Detailed task- and rank-wise results, descriptive intervals, and win/tie/loss summaries are provided in Section D.

We also evaluate WAS-LoRA on Qwen2.5-1.5B-Instruct (Yang et al., 2024) with GSM8K and ARC-Challenge. The aggregate fidelity advantage persists under both matched-total-rank and physical-parameter-matched comparisons (Section F).

4.2 Does fidelity translate into ground-truth performance?

Figure 3 shows the accuracy of WAS-LoRA and the baseline methods across five average retained ranks ranging from 2 to 8. Relative to Act-SVD, the five-task mean accuracy gains of WAS-LoRA are $+0.4600$, $+0.2213$, $+0.1294$, $+0.0566$, and -0.0312 percentage points at average retained ranks $\bar{r} \in \{2, 3, 4, 6, 8\}$, respectively. Against AdaLoRA-score, the corresponding gains are $+0.3597$, $+0.1998$, $+0.0688$, -0.0047, and -0.0635 percentage points at average retained ranks $\bar{r} \in \{2, 3, 4, 6, 8\}$, respectively. Complete paired estimates, descriptive intervals, and win/tie/loss counts are reported in Section D. From the results shown in Figure 3, we can conclude that WAS-LoRA achieves its largest mean accuracy gains under tight retained-rank budgets, while maintaining mean accuracy comparable to that of the baselines at higher ranks.

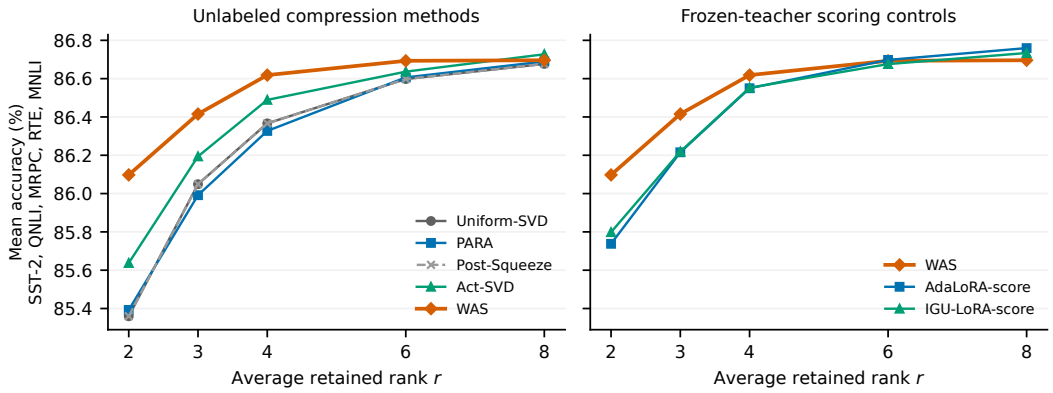


Figure 3: **Accuracy benefits depend on the rank budget.** Ordinary classification accuracy, averaged over eight paired seeds within each task and then equally over the five tasks. MRPC uses raw accuracy; MNLI averages matched/mismatched accuracy within each teacher cell. The right panel contains frozen-teacher scoring controls. Uniform-SVD and Post-Squeeze nearly overlap. Paired uncertainty is reported in the appendix.

4.3 Does WAS-LoRA improve on fixed-basis rescoring?

Output-scored Act-SVD incorporates downstream sensitivity while keeping the local basis fixed, reducing geometric-mean centered relative logit error by 17.0% relative to Act-SVD. WAS-LoRA allows downstream sensitivity to construct new directions and achieves a further 15.8% reduction relative to Output-scored Act-SVD.

In a complementary held-out audit on three SST-2 teachers, we restrict both criteria to the same 384 physical rank-one SVD candidates. The downstream-sensitivity score used by WAS-LoRA correlates more strongly with their exact single-deletion effects than local singular energy (Spearman $\rho = 0.955$–0.959 versus 0.857–0.863; Figure 5).

4.4 Sensitivity to calibration-set size

We vary the MNLI calibration size over $N \in \{32, 64, 128, 256, 2048\}$ at the budget $K = 192$, using three frozen teachers and five fixed nested draws. As shown in Figure 4, even $N = 32$ (1.56% of the full pool) changes mean accuracy by only 0.044 percentage points. Across all tested sizes, the largest absolute mean changes in centered teacher-logit error and KL are 7.4×10^{-4} and 1.2×10^{-4}, respectively. These results indicate that the final compressed model is insensitive to substantial calibration subsampling in this setting. Cross-task geometry curves, complete values, and the tighter $K = 48$ case are reported in Section C.3.

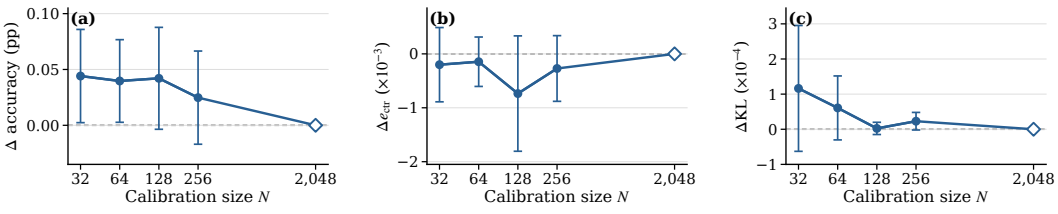


Figure 4: **WAS-LoRA is insensitive to calibration-set size on MNLI at $K = 192$.** Points are mean paired changes from the $N = 2048$ full-pool reference. Five fixed draws are averaged within each of three teachers; error bars show standard deviation across the teacher means. Accuracy and both teacher-fidelity metrics remain close to the reference throughout the measured range.

4.5 COMPUTATIONAL COST

In our implementation, we measure the computational cost on a single NVIDIA A800-SXM4-80GB GPU. Collecting calibration statistics takes 1.91 seconds for Act-SVD, 5.95 seconds for WAS-LoRA, 46.46 seconds for AdaLoRA-score, and 89.39 seconds for IGU-LoRA-score. These statistics are collected once per teacher and reused across retained-rank budgets. Factor construction adds negligible overhead, allowing a single teacher to support multiple compressed variants without repeated calibration or recovery training. The reported timings are implementation-specific and concern offline compression, not inference latency. Aggregate and per-task timings, memory usage, and measurement details are provided in Section E.5.

5 RELATED WORK

LoRA serving and post-training compression. Multi-adapter serving motivates reducing the deployment cost of trained LoRAs (Sheng et al., 2024; Chen et al., 2024; Wu et al., 2024; Li et al., 2025b; Gabrielsson et al., 2025). Spectral methods extract LoRAs from full-rank checkpoints (Vasani et al., 2025) or compress trained LoRAs (Vulić et al., 2026; Kumaravelu et al., 2026); Compress then Merge compresses before merging (He et al., 2026). WAS-LoRA uses unlabeled calibration inputs without recovery training, a practical setting shared with broader compression methods such as AdaRound and SparseGPT (Nagel et al., 2020; Frantar and Alistarh, 2023).

Adaptive rank and sensitivity-based LoRA. LoRA-drop uses continued optimization, while NoRA combines outer-factor initialization with nested fine-tuning (Zhou et al., 2025; Li et al., 2025a). AdaLoRA, IGU-LoRA, and GoRA use gradient-derived signals for rank allocation or initialization (Zhang et al., 2023; Cui et al., 2026; He et al., 2025). Spectral Surgery reweights singular values using calibration-loss gradients while retaining the singular directions (Tian et al., 2026). In the fixed-teacher setting, WAS-LoRA uses sensitivity to a selected teacher output to construct a retained subspace, allowing new linear combinations of whitened bottleneck coordinates rather than restricting selection to existing components.

Activation- and output-aware low-rank approximation. For dense-model compression, ASVD and SVD-LLM use activation statistics to guide low-rank approximation (Yuan et al., 2024; Wang et al., 2025b), while Dobi-SVD optimizes activation truncation and weight reconstruction (Wang et al., 2025a). IO-SVD incorporates a predictive-KL output metric through double-sided whitening (Abbasi et al., 2026). WAS-LoRA likewise uses input and output information, but operates within a trained LoRA bottleneck rather than a full dense weight. Its ridge-free $H = 0$ case recovers the exact local activation-aware optimum, while the main experiments use Euclidean raw-logit sensitivity for downstream output preservation.

Active Subspaces. Our sensitivity criterion builds on derivative-based Active Subspaces (Constantine et al., 2014; Constantine and Gleich, 2015), vector-valued extensions (Zahm et al., 2020), and neural-network applications such as ASNet (Cui et al., 2020).

6 CONCLUSION

We introduced WAS-LoRA, a post-training LoRA compression method that uses the sensitivity of a teacher output to choose a retained subspace in whitened bottleneck coordinates. It constructs standard lower-rank LoRA factors from unlabeled calibration inputs without recovery training. At $H = 0$, the ridge-free construction recovers the exact local activation-aware rank-k optimum. Experiments show that WAS-LoRA achieves teacher-function fidelity gains. Task-accuracy differences are concentrated at tight retained-rank budgets. A central theoretical **limitation** is that our guarantees concern first-order discarded sensitivity, with global rank allocation additionally relying on a module-wise additive objective. They do not establish optimality for joint finite nonlinear compression. Important directions are to connect this surrogate to finite compression error and to adapt the propagation horizon across modules and budgets. The same geometry may also inform training-time rank allocation and compression of other learned bottlenecks.

AI USE STATEMENT

The authors defined the research questions, methodological direction, and final scientific claims. Generative AI tools were used in an assistive role for methodological discussion, experimental analysis, manuscript editing, and translation. All mathematical claims and proofs were checked by the authors; AI-assisted code was inspected, executed, and tested; experimental results and citations were independently verified by the authors. The authors made all final scientific and editorial decisions and take full responsibility for the content of the submission.

REPRODUCIBILITY STATEMENT

Proofs and assumptions for the main theoretical results are provided in Sections A.1–A.5. Section B gives analytic examples; Sections C.1–C.3 report the subspace-selection analyses; and Section D provides the full RoBERTa/GLUE results. Section E documents the experimental protocol, implementation, and measured costs, while Section E.6 identifies the verification records and regeneration scripts. Section F provides the Qwen2.5-1.5B-Instruct experimental protocol and results.

REFERENCES

A. Abbasi, C. Thrash, H. Qin, H. Pirsiavash, and S. Kolouri. IO-SVD: Input-output whitened SVD for adaptive-rank LLM compression. *arXiv:2605.15626*, 2026.

L. Chen, Z. Ye, Y. Wu, D. Zhuo, L. Ceze, and A. Krishnamurthy. Punica: Multi-tenant LoRA serving. In *MLSys*, 2024.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457*, 2018.

K. Cobbe, V. Kosaraju, M. Bavarian, et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

P. G. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4), 2014.

P. G. Constantine and D. F. Gleich. Computing active subspaces with Monte Carlo. *arXiv:1408.0545v2*, 2015.

C. Cui, K. Zhang, T. Daulbaev, J. Gusak, I. Oseledets, and Z. Zhang. Active subspace of neural networks: Structural analysis and universal attacks. *SIAM Journal on Mathematics of Data Science*, 2(4):1096–1122, 2020. DOI: 10.1137/19M1296070.

X. Cui, H. Li, R. Zeng, et al. IGU-LoRA: Adaptive rank allocation via integrated gradients and uncertainty-aware scoring. In *ICLR*, 2026.

E. Frantar and D. Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *ICML*, pp. 10323–10337, 2023.

R. B. Gabrielsson, J. Zhu, O. Bhardwaj, L. Choshen, K. Greenewald, M. Yurochkin, and J. Solomon. Compress then Serve: Serving thousands of LoRA adapters with little overhead. In *ICML*, 2025.

H. He, P. Ye, Y. Ren, Y. Yuan, L. Zhou, S. Ju, and L. Chen. GoRA: Gradient-driven adaptive low rank adaptation. In *NeurIPS*, 2025.

Z. He, R. Ding, Z. Huang, R. Yang, T. Li, and X. Huang. Compress then Merge: From multiple LoRAs into one low-rank adapter. In *ICML*, 2026.

E. J. Hu, Y. Shen, P. Wallis, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 19(2), 1990.

V. Kumaravelu, S. Gupta, and P. K. Srijith. Post-optimization adaptive rank allocation for LoRA. *arXiv:2604.27796*, 2026.

L. Li, C. Lin, D. Li, Y.-L. Huang, W. Li, T. Wu, J. Zou, W. Xue, S. Han, and Y. Guo. Efficient fine-tuning of large models via nested low-rank adaptation. In *ICCV*, pp. 22252–22262, 2025.

S. Li, H. Lu, T. Wu, M. Yu, Q. Weng, X. Chen, Y. Shan, B. Yuan, and W. Wang. Toppings: CPU-assisted, rank-aware adapter serving for LLM inference. In *USENIX ATC*, pp. 613–629, 2025.

Y. Liu, M. Ott, N. Goyal, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.

M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *ICML*, pp. 7197–7206, 2020.

Y. Sheng, S. Cao, D. Li, C. Hooper, N. Lee, S. Yang, C. Chou, B. Zhu, L. Zheng, K. Keutzer, J. E. Gonzalez, and I. Stoica. S-LoRA: Serving thousands of concurrent LoRA adapters. In *MLSys*, 2024.

Z. Tian, Y. Chen, Z. Han, and L. Liao. Spectral surgery: Training-free refinement of LoRA via gradient-guided singular value reweighting. *arXiv:2603.03995*, 2026.

M. Valipour, M. Rezagholizadeh, I. Kobyzev, and A. Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *EACL*, pp. 3274–3287, 2023.

B. Vasani, J. FitzGerald, A. Fang, and S. Vaish. PHLoRA: Data-free post-hoc low-rank adapter extraction from full-rank checkpoint. In *Findings of IJCNLP-AACL*, 2025.

I. Vulić, A. Grycner, Q. de Laroussilhe, and J. Pfeiffer. LoRA-Squeeze: Simple and effective post-tuning and in-tuning compression of LoRA modules. *arXiv:2602.10993*, 2026.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP*, 2018.

Q. Wang, J. Ke, M. Tomizuka, K. Keutzer, and C. Xu. Dobi-SVD: Differentiable SVD for LLM compression and some new perspectives. In *ICLR*, 2025.

X. Wang, Y. Zheng, Z. Wan, and M. Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. In *ICLR*, 2025.

B. Wu, R. Zhu, Z. Zhang, P. Sun, X. Liu, and X. Jin. dLoRA: Dynamically orchestrating requests and adapters for LoRA LLM serving. In *OSDI*, pp. 911–927, 2024.

A. Yang, B. Yang, B. Zhang, et al. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.

Z. Yuan, Y. Shang, Y. Song, et al. ASVD: Activation-aware singular value decomposition for compressing large language models. *arXiv:2312.05821v4*, 2024.

O. Zahm, P. G. Constantine, C. Prieur, and Y. Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1), 2020.

Q. Zhang, M. Chen, A. Bukharin, et al. Adaptive budget allocation for parameter-efficient fine-tuning. In *ICLR*, 2023.

H. Zhou, X. Lu, W. Xu, et al. LoRA-drop: Efficient LoRA parameter pruning based on output evaluation. In *COLING*, 2025.

A Proofs

A.1 Proof of Theorem 1: Optimal discarded downstream sensitivity

Since P is an orthogonal projector, we have $P = P^\top = P^2$. Let P have rank k. Expanding the squared Frobenius norm and using $\|A\|_F^2 = \mathrm{tr}(A^\top A)$ gives

$$R_H^{\mathrm{w}}(P) = \mathbb{E}_x \|\widetilde{J}_H(x)(I-P)\|_F^2 \tag{9}$$

$$= \mathbb{E}_x \,\mathrm{tr}\Big((I-P)\widetilde{J}_H(x)^\top \widetilde{J}_H(x)(I-P)\Big) \tag{10}$$

$$= \mathrm{tr}\big((I-P)M_H(I-P)\big)\,. \tag{11}$$

Since P is an orthogonal projector, $(I-P)^2 = I-P$, and by cyclicity of the trace,

$$R_H^{\mathrm{w}}(P) = \mathrm{tr}((I-P)M_H) = \mathrm{tr}(M_H) - \mathrm{tr}(PM_H). \tag{12}$$

Therefore minimizing the discarded sensitivity is equivalent to maximizing $\mathrm{tr}(PM_H)$.

Let $M_H = U\Lambda U^\top$ be the eigendecomposition, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r$. Define $\alpha_i = u_i^\top P u_i$. Because P is an orthogonal projector, $0 \leq \alpha_i \leq 1$. Moreover,

$$\sum_i \alpha_i = \sum_i u_i^\top P u_i = \mathrm{tr}(P) = k.$$

Hence

$$\mathrm{tr}(PM_H) = \sum_i \lambda_i \alpha_i. \tag{13}$$

The maximum is obtained by assigning the available unit weights to the largest eigenvalues, namely $\alpha_1 = \cdots = \alpha_k = 1, \quad \alpha_{k+1} = \cdots = \alpha_r = 0$. Equivalently, by the Ky Fan maximum principle,

$$\max_{\substack{P=P^\top=P^2 \\ \mathrm{rank}(P)=k}} \mathrm{tr}(PM_H) = \sum_{i=1}^{k} \lambda_i(M_H), \tag{14}$$

attained by $P = U_k U_k^\top$, with the usual nonuniqueness inside tied eigenspaces. Subtracting this maximum from $\mathrm{tr}(M_H)$ proves Theorem 1.

A.2 Exact global rank allocation under equal direction costs

We now show that the global eigenvalue pooling step in Algorithm 1 is not merely a heuristic: once the module-wise sensitivity matrices are fixed, it exactly solves the additive discarded-sensitivity problem under a total rank budget when every retained direction has the same cost.

For each module ℓ, let

$$M_{H,\ell} = U_\ell \Lambda_\ell U_\ell^\top \succeq 0,$$

where

$$\lambda_{\ell,1} \geq \lambda_{\ell,2} \geq \cdots \geq \lambda_{\ell,d_\ell} \geq 0$$

are the eigenvalues on its admissible coordinate space of dimension d_ℓ. In the ridge-free construction, d_ℓ may be taken as the dimension of the data-supported subspace. Let m_ℓ denote the required minimum retained rank of module ℓ.

Proposition 4 (Exact global allocation under equal direction costs)**.** *Assume that*

$$0 \leq m_\ell \leq d_\ell$$

for every module ℓ, and let the total rank budget K be feasible:

$$\sum_\ell m_\ell \leq K \leq \sum_\ell d_\ell.$$

Consider the optimization problem

$$\begin{aligned} \min_{\{P_\ell\}} \quad & \sum_\ell \mathrm{tr}((I - P_\ell) M_{H,\ell}) \\ \textit{subject to} \quad & P_\ell = P_\ell^\top = P_\ell^2, \\ & m_\ell \leq \mathrm{rank}(P_\ell) \leq d_\ell, \\ & \sum_\ell \mathrm{rank}(P_\ell) = K. \end{aligned} \tag{15}$$

After reserving the leading m_ℓ directions of every module, pool the remaining eligible eigenvalues

$$\mathcal{E} = \{\lambda_{\ell,j} : m_\ell < j \leq d_\ell\}_\ell .$$

Let

$$q = K - \sum_\ell m_\ell.$$

Selecting the q largest values in $\mathcal{E}$ and assigning each selected value to its corresponding module yields a globally optimal rank allocation. More precisely, if $k_\ell^\star$ is the resulting rank of module ℓ, then

$$P_\ell^\star = U_{\ell,k_\ell^\star} U_{\ell,k_\ell^\star}^\top$$

solves the optimization problem defined in Eq. (15)*.*

Ties at the global cutoff are resolved to preserve the within-module prefix ordering. All resulting selections attain the same optimal objective value.

Proof. We first fix a feasible collection of module ranks $\{k_\ell\}$ satisfying

$$m_\ell \le k_\ell \le d_\ell, \qquad \sum_\ell k_\ell = K.$$

By Theorem 1, for each individual module ℓ,

$$\min_{\substack{P_\ell = P_\ell^\top = P_\ell^2 \\ \mathrm{rank}(P_\ell) = k_\ell}} \mathrm{tr}((I - P_\ell) M_{H,\ell}) = \sum_{j > k_\ell} \lambda_{\ell,j}, \tag{16}$$

and the minimum is attained by the leading-eigenspace projector

$$P_\ell = U_{\ell,k_\ell} U_{\ell,k_\ell}^\top.$$

Therefore, once the ranks $\{k_\ell\}$ are fixed, the module-wise subspace choices are independent and jointly attain

$$\sum_\ell \sum_{j > k_\ell} \lambda_{\ell,j}. \tag{17}$$

Hence the only remaining problem is the discrete allocation of the ranks k_ℓ. Since

$$\sum_{j > k_\ell} \lambda_{\ell,j} = \sum_{j=1}^{d_\ell} \lambda_{\ell,j} - \sum_{j=1}^{k_\ell} \lambda_{\ell,j},$$

minimizing (17) is equivalent to maximizing

$$\sum_\ell \sum_{j=1}^{k_\ell} \lambda_{\ell,j}. \tag{18}$$

The contribution of the mandatory first m_ℓ directions is fixed, so the variable part of (18) is

$$\sum_\ell \sum_{j=m_\ell+1}^{k_\ell} \lambda_{\ell,j}, \tag{19}$$

subject to

$$\sum_\ell (k_\ell - m_\ell) = K - \sum_\ell m_\ell = q.$$

Thus every additional rank slot retains one eligible eigenvalue, and its exact marginal reduction in discarded sensitivity is that eigenvalue. In particular, increasing module ℓ from rank k_ℓ to $k_\ell + 1$ changes its optimal residual from

$$\sum_{j > k_\ell} \lambda_{\ell,j} \quad \textit{to} \quad \sum_{j > k_\ell + 1} \lambda_{\ell,j},$$

and therefore reduces the objective by exactly $\lambda_{\ell,k_\ell+1}$.

It remains to determine which q eligible marginal gains should be selected. Any feasible allocation $\{k_\ell\}$ corresponds to the set

$$S(\{k_\ell\}) = \{(\ell, j) : m_\ell < j \le k_\ell\},$$

which contains exactly q eligible eigenvalues. Consequently,

$$\sum_\ell \sum_{j=m_\ell+1}^{k_\ell} \lambda_{\ell,j} = \sum_{(\ell,j) \in S(\{k_\ell\})} \lambda_{\ell,j}. \tag{20}$$

Among all subsets of q eligible values, the largest possible sum is obtained by choosing the globally largest q eigenvalues. Moreover, because the spectrum of every module is nonincreasing,

$$\lambda_{\ell,j-1} \geq \lambda_{\ell,j},$$

this global selection can always be chosen to respect the within-module prefix constraint: if a selected direction (ℓ, j) appears without an eligible predecessor $(\ell, j-1)$, replacing (ℓ, j) by $(\ell, j-1)$ cannot decrease the retained sum. Repeating this replacement yields a prefix-closed selection of the same cardinality and with no smaller objective. When equality occurs, this is simply a tie-breaking choice.

Let $S^\star$ be such a prefix-closed set of the globally largest q eligible eigenvalues, and define

$$k_\ell^\star = m_\ell + \#\left\{j : (\ell, j) \in S^\star\right\}.$$

Then $\{k_\ell^\star\}$ is feasible, and for every other feasible allocation $\{k_\ell\}$,

$$\sum_{(\ell,j)\in S(\{k_\ell\})} \lambda_{\ell,j} \leq \sum_{(\ell,j)\in S^\star} \lambda_{\ell,j}.$$

Therefore $\{k_\ell^\star\}$ maximizes (19), equivalently minimizes (17). Combining this allocation with the module-wise leading-eigenspace projectors from Theorem 1 proves (15). □

Scope. Proposition 4 is exact for the additive discarded-sensitivity surrogate with fixed module matrices $M_{H,\ell}$ and equal per-direction costs. It does not assert global optimality for joint finite nonlinear compression across modules. Moreover, if retained directions have unequal physical costs, the allocation problem becomes a budgeted discrete optimization problem rather than simple global eigenvalue sorting.

A.3 LoRA-compatible reconstruction

For a raw-coordinate retained basis $V \in \mathbb{R}^{r\times k}$, let $z = V^\top s$, $A' = V^\top A$, and define $G = \mathbb{E}[ss^\top]$. The compressed LoRA branch has the form $B'V^\top s$. Therefore the optimal decoder is obtained from the linear least-squares problem

$$B'_\star \in \arg\min_{B'} \mathbb{E}\|Bs - B'V^\top s\|_2^2. \tag{21}$$

Expanding the objective gives

$$\begin{aligned}\mathcal{L}(B') &= \mathbb{E}\,\mathrm{tr}((Bs - B'V^\top s)(Bs - B'V^\top s)^\top) \\ &= \mathrm{tr}(BGB^\top) - 2\,\mathrm{tr}(B'V^\top GB^\top) + \mathrm{tr}(B'V^\top GVB'^\top).\end{aligned} \tag{22}$$

Let $G_V = V^\top GV \succeq 0$. The first-order optimality condition of this convex quadratic objective is

$$B'G_V = BGV. \tag{23}$$

We first verify that this equation is consistent even when G_V is singular. Suppose $G_V a = 0$. Then $0 = a^\top G_V a = a^\top V^\top GVa = (Va)^\top G(Va)$. Since $G = \mathbb{E}[ss^\top]$, $(Va)^\top G(Va) = \mathbb{E}[((Va)^\top s)^2]$. Therefore $(Va)^\top s = 0$ almost surely, which implies $GVa = \mathbb{E}[ss^\top Va] = 0$. Hence every vector in the null space of G_V is also in the null space of GV, proving that (23) is solvable.

The minimum-norm solution of this consistent linear system is

$$B'_\star = BGV(V^\top GV)^\dagger. \tag{24}$$

For standardized coordinates, $G = I$, and for an orthonormal basis V, $V^\top GV = V^\top V = I_k$. Therefore $B'_\star = BV$, which gives the absorbed LoRA factors in (7).

A.4 ZERO-HORIZON EQUIVALENCE

Let the compact SVD of the observed bottleneck activations be

$$AX = L_s \Gamma Q^\top, \qquad D = \gamma B L_s \Gamma, \tag{25}$$

where Γ contains the positive singular values on the data-supported subspace. Since the calibration matrix X is scaled as in Theorem 2, the empirical second moment of the bottleneck activation is

$$\Sigma = (AX)(AX)^\top = L_s \Gamma^2 L_s^\top. \tag{26}$$

Hence, on the data-supported subspace,

$$\Sigma^{1/2} = L_s \Gamma L_s^\top, \qquad \Sigma^{-1/2} = L_s \Gamma^{-1} L_s^\top. \tag{27}$$

We first connect the WAS construction to the SVD above. At $H = 0$, the selected endpoint is the immediate LoRA branch output, $q_0(s) = \gamma B s$, so its Jacobian with respect to the bottleneck activation s is $J_0 = \gamma B$. Therefore the whitened Jacobian and its Gram matrix are

$$\widetilde{J}_0 = J_0 \Sigma^{1/2} = \gamma B L_s \Gamma L_s^\top, \tag{28}$$

$$M_0 = \widetilde{J}_0^\top \widetilde{J}_0 \tag{29}$$

$$= L_s \left(\Gamma L_s^\top \gamma^2 B^\top B L_s \Gamma \right) L_s^\top \tag{30}$$

$$= L_s D^\top D L_s^\top. \tag{31}$$

Let V_k contain the leading right singular vectors of D. Since the columns of V_k are the leading eigenvectors of $D^\top D$, Equation (31) shows that the leading eigenspace selected by WAS is

$$U_k = L_s V_k. \tag{32}$$

Substituting Equations (27) and (32) into the WAS absorption rule gives the physical compressed update

$$\gamma B_k A_k = \gamma B \Sigma^{1/2} U_k U_k^\top \Sigma^{-1/2} A \tag{33}$$

$$= \gamma B L_s \Gamma V_k V_k^\top \Gamma^{-1} L_s^\top A. \tag{34}$$

Thus the update produced by ridge-free $H = 0$ WAS is precisely

$$W_k = \gamma B L_s \Gamma V_k V_k^\top \Gamma^{-1} L_s^\top A. \tag{35}$$

It remains to identify the output induced by this update and prove its activation-aware optimality. The immediate LoRA branch output on the calibration activations is $\gamma BAX = \gamma B L_s \Gamma Q^\top = D Q^\top$. Because $Q^\top Q = I$ on the data-supported subspace, right multiplication by $Q^\top$ preserves the nonzero singular values. More explicitly, if $D v_i = \sigma_i u_i$, then $\gamma BAX(Q v_i) = D Q^\top Q v_i = D v_i = \sigma_i u_i$. Hence the right singular vectors of γBAX on the supported subspace are obtained by mapping those of D through Q.

Multiplying Equation (35) by X gives

$$W_k X = \gamma B L_s \Gamma V_k V_k^\top \Gamma^{-1} L_s^\top A X \tag{36}$$

$$= \gamma B L_s \Gamma V_k V_k^\top Q^\top \tag{37}$$

$$= D V_k V_k^\top Q^\top. \tag{38}$$

The last expression is exactly the rank-k truncated SVD of γBAX:

$$(\gamma BAX)_k = D V_k V_k^\top Q^\top. \tag{39}$$

It remains to show optimality among all feasible physical updates. For any matrix $\widetilde{W}$ satisfying

$$\operatorname{rank}(\widetilde{W}) \le k,$$

we have

$$\operatorname{rank}(\widetilde{W} X) \le k.$$

Therefore every feasible compressed LoRA update induces a rank-at-most-k approximation of the observed output matrix γBAX.

By the Eckart–Young–Mirsky theorem, the truncated SVD $(\gamma BAX)_k$ is the optimal rank-k approximation under the Frobenius norm. Since the $H = 0$ WAS construction above attains this truncated SVD exactly,

$$W_k \in \arg \min_{\mathrm{rank}(\widetilde{W}) \le k} \|(\gamma BA - \widetilde{W})X\|_F^2.$$

If AX is rank deficient, all square roots and inverse square roots above are understood on the data-supported subspace; equivalently, Γ^{-1} is replaced by the Moore–Penrose pseudoinverse on that subspace. This proves Theorem 2.

A.5 Proof of Proposition 3: Gauge invariance

For any invertible matrix $R \in \mathbb{R}^{r \times r}$, the refactorization

$$A' = RA, \qquad B' = BR^{-1} \tag{40}$$

represents the same LoRA update. Let $s = Ah$ and $s' = A'h = Rs$. For the same calibration distribution and the same output q_H, define

$$\Sigma = \mathbb{E}[ss^\top], \qquad C_H = \mathbb{E}[J_H(x)^\top J_H(x)],$$

where $J_H(x)$ is the Jacobian of q_H with respect to s. By the chain rule,

$$\Sigma' = R\Sigma R^\top, \qquad J'_H(x) = J_H(x)R^{-1}, \qquad C'_H = R^{-\top} C_H R^{-1}.$$

Let $d = \mathrm{rank}(\Sigma)$. Since R is invertible, $\mathrm{rank}(\Sigma') = d$. Write the compact eigendecompositions

$$\Sigma = L\Lambda L^\top, \qquad \Sigma' = L'\Lambda' L'^\top,$$

where $L, L' \in \mathbb{R}^{r \times d}$ have orthonormal columns and $\Lambda, \Lambda' \succ 0$. Define the compact square-root factors

$$F = L\Lambda^{1/2}, \qquad F' = L'\Lambda'^{1/2},$$

so that $FF^\top = \Sigma$ and $F'F'^\top = \Sigma'$, and define

$$O = F'^{\dagger} RF \in \mathbb{R}^{d \times d}.$$

Because

$$\Sigma' = RFF^\top R^\top = (RF)(RF)^\top,$$

we have

$$\mathrm{range}(RF) = \mathrm{range}(\Sigma') = \mathrm{range}(F').$$

Hence

$$RF = F'O.$$

Moreover,

$$\begin{aligned} OO^\top &= F'^{\dagger} RFF^\top R^\top (F'^{\dagger})^\top \\ &= F'^{\dagger} \Sigma' (F'^{\dagger})^\top \\ &= I_d. \end{aligned} \tag{41}$$

Since O is square, it is orthogonal:

$$O^\top O = OO^\top = I_d.$$

Equivalently,

$$R^{-1}F' = FO^\top. \tag{42}$$

Define the compact sensitivity matrix

$$M_{\mathrm{c}} = F^\top C_H F.$$

Because $\Sigma^{1/2} = L\Lambda^{1/2}L^\top$, the ambient sensitivity matrix satisfies

$$M_H = \Sigma^{1/2} C_H \Sigma^{1/2} = L M_{\mathrm{c}} L^\top.$$

Similarly,

$$M'_H = L' M'_{\mathrm{c}} L'^\top,$$

where, using (42),

$$\begin{aligned} M'_{\mathrm{c}} &= F'^\top C'_H F' \\ &= F'^\top R^{-\top} C_H R^{-1} F' \\ &= O M_{\mathrm{c}} O^\top. \end{aligned} \tag{43}$$

Thus M_{c} and M'_{c} are orthogonally similar and have identical supported eigenvalues.

For $0 \le k \le d$, let $V_k \in \mathbb{R}^{d\times k}$ contain a set of leading eigenvectors of M_{c}, and choose the corresponding eigenvectors

$$V'_k = OV_k$$

for M'_{c}. Their projectors satisfy

$$P'_k = OP_kO^\top, \qquad P_k = V_kV_k^\top.$$

If the cutoff splits a tied eigenspace, the retained subspaces must be matched in this way; arbitrary independent choices inside the tied eigenspace need not yield identical compressed outputs. When there is a spectral gap at the cutoff, the leading projector is unique.

On the data-supported subspace, the compressed-factor construction can be written as

$$A_k = V_k^\top F^\dagger A, \qquad B_k = BFV_k,$$

because the corresponding ambient eigenvectors are $U_k = LV_k$. For the refactorized adapter,

$$\begin{aligned} B'_k &= B'F'V'_k \\ &= BR^{-1}F'OV_k \\ &= BFO^\top OV_k \\ &= BFV_k \\ &= B_k. \end{aligned} \tag{44}$$

Now consider any input for which

$$s = Ah \in \mathrm{range}(\Sigma) = \mathrm{range}(F).$$

Writing $s = Fv$, with $v = F^\dagger s$, gives

$$F'^\dagger Rs = F'^\dagger RFv = Ov = OF^\dagger s.$$

Therefore

$$\begin{aligned} A'_k h &= V_k'^\top F'^\dagger A' h \\ &= V_k^\top O^\top F'^\dagger RAh \\ &= V_k^\top F^\dagger Ah \\ &= A_k h. \end{aligned} \tag{45}$$

Consequently,

$$B'_k A'_k h = B_k A_k h \qquad \text{whenever } Ah \in \mathrm{range}(\Sigma).$$

The common LoRA scale γ preserves this equality.

In particular, since $\Sigma = \mathbb{E}[ss^\top]$, calibration-distributed bottleneck activations satisfy $s \in \mathrm{range}(\Sigma)$ almost surely. Hence ridge-free WAS-LoRA is gauge invariant on the data-supported subspace. If $\Sigma \succ 0$, then $\mathrm{range}(\Sigma) = \mathbb{R}^r$, so the equality holds for every input and

$$B'_k A'_k = B_k A_k.$$

For singular Σ, equality outside the data-supported subspace is not guaranteed. If $d = 0$, the only supported compressed update is the zero update.

This proves Proposition 3.

B Analytic Examples

B.1 Finite-loss verification for the LoRA–ReLU example

We first verify the finite-loss calculation in Section 2.1 and then connect the same example explicitly to the WAS-LoRA construction. The frozen backbone and bias are $W_0 = I_2, \quad b = (-20, 3)^\top$, and the trained rank-two LoRA update is $\Delta W = BA = \mathrm{diag}(10, 1), \quad A = I_2, \quad B = \mathrm{diag}(10, 1)$. For $x \sim \mathrm{Unif}([-1, 1]^2)$, the teacher is $f(x) = \mathbf{1}^\top \mathrm{ReLU}\big((W_0 + \Delta W)x + b\big) = 3 + 2x_2$, since $11x_1 - 20 < 0, \quad 2x_2 + 3 > 0$ throughout the support.

Local low-rank compression. Weight SVD searches over arbitrary rank-one matrices and gives the unique optimum

$$\widetilde{W}_{\mathrm{SVD}} = \mathrm{diag}(10, 0).$$

The activation-aware local objective yields the same solution, because for every candidate matrix C,

$$\mathbb{E}\|(\Delta W - C)x\|_2^2 = \frac{1}{3}\|\Delta W - C\|_F^2.$$

The resulting output is therefore $f_{\mathrm{SVD}}(x) = 3 + x_2$, and hence

$$\mathcal{L}(\widetilde{W}_{\mathrm{SVD}}) = \mathbb{E}\big|f(x) - f_{\mathrm{SVD}}(x)\big|^2 = \mathbb{E}[x_2^2] = \frac{1}{3}.$$

Downstream sensitivity. Now consider the same trained update through the bottleneck parameterization used by WAS-LoRA. The bottleneck variable and its uncentered second moment are $s = Ax = x, \quad \Sigma = \mathbb{E}[ss^\top] = \frac{1}{3} I_2, \quad u = \Sigma^{-1/2} s = \sqrt{3}\, x$. To differentiate only the adapter intervention while holding the frozen path fixed, write $F(x, u) = \mathbf{1}^\top \mathrm{ReLU}\big(W_0 x + B\Sigma^{1/2} u + b\big)$. At the observed teacher bottleneck $u = \sqrt{3}x$, the downstream gates are fixed as above, and therefore $J_u = \frac{\partial F(x,u)}{\partial u} = \frac{1}{\sqrt{3}} e_2^\top$. The whitened sensitivity matrix is $M = \mathbb{E}[J_u^\top J_u] = \frac{1}{3} e_2 e_2^\top$. Its unique

leading eigenvector is e_2. Absorbing this direction into ordinary rank-one LoRA factors gives $B_1 = B\Sigma^{1/2}e_2 = \frac{1}{\sqrt{3}}e_2, \quad A_1 = e_2^\top \Sigma^{-1/2} A = \sqrt{3}\, e_2^\top$, so that $B_1 A_1 = e_2 e_2^\top = \mathrm{diag}(0,1)$. This compressed adapter exactly preserves the teacher output,

$$f_{\mathrm{WAS}}(x) = 3 + 2x_2 = f(x), \qquad \mathcal{L}(\widetilde{W}_{\mathrm{WAS}}) = 0.$$

The agreement is even stronger in this example. For any rank-one orthogonal projector P, isotropic whitening gives the compressed update BPA. Moreover, for $x \in [-1,1]^2$, $|(Px)_i| \le \sqrt{2}$, so the first downstream gate remains inactive and the second remains active. Hence $f_P(x) = 3 + x_2 + e_2^\top Px$ and $\mathcal{L}(P) = \frac{1}{3} e_2^\top (I-P) e_2$. At the same time,

$$R^{\mathrm{w}}(P) = \mathbb{E}\|J_u(I-P)\|_F^2 = \frac{1}{3} e_2^\top (I-P) e_2.$$

Thus in this construction the discarded-sensitivity criterion agrees exactly with the finite output loss over the entire rank-one projector family.

B.2 A SMOOTH FAMILY OF LOCAL-TO-DOWNSTREAM REVERSALS

The ReLU construction in Section B.1 uses an exactly closed downstream gate. The same reversal occurs for a smooth nonlinearity without hard gating.

Let $\phi(z) = \tanh(z)$ act componentwise and consider the vector-valued teacher endpoint

$$F_{\widetilde{W}}(x) = \tanh\big((W_0 + \widetilde{W})x + b\big), \qquad W_0 = I_2, \qquad \Delta W = \mathrm{diag}(a,c), \qquad b = (\beta, 0)^\top, \tag{46}$$

where $a > c > 0, \quad x \sim \mathrm{Unif}([-1,1]^2)$. Since $\mathbb{E}[xx^\top] = I_2/3$, both weight SVD and activation-aware local reconstruction retain the larger first direction,

$$\widetilde{W}_{\mathrm{SVD}} = \mathrm{diag}(a, 0).$$

The finite endpoint error of this compression is

$$\mathcal{L}_{\mathrm{SVD}} = \mathbb{E}\left[\tanh\big((1+c)x_2\big) - \tanh(x_2)\right]^2 =: \delta_c > 0. \tag{47}$$

Now consider the WAS-LoRA sensitivity. Using the LoRA factorization $A = I_2, \quad B = \mathrm{diag}(a,c)$, the bottleneck uncentered second moment is $\Sigma = I_2/3$. For the whitened coordinate $u = \Sigma^{-1/2}s$, the endpoint Jacobian at the teacher is

$$J_u = \frac{1}{\sqrt{3}} \mathrm{diag}\big(a\ \mathrm{sech}^2\big(\beta + (1+a)x_1\big), c\ \mathrm{sech}^2\big((1+c)x_2\big)\big)\,.$$

Hence

$$M = \frac{1}{3} \mathrm{diag}\big(a^2 \mathbb{E}\,\mathrm{sech}^4\big(\beta + (1+a)x_1\big), c^2 \mathbb{E}\,\mathrm{sech}^4\big((1+c)x_2\big)\big)\,. \tag{48}$$

The second eigenvalue is a fixed positive constant, whereas the first satisfies

$$\frac{a^2}{3} \mathbb{E}\,\mathrm{sech}^4\big(\beta + (1+a)x_1\big) \longrightarrow 0 \qquad \text{as } \beta \to \infty.$$

Therefore, for every $a > c > 0$, there exists a finite $\beta_{\mathrm{WAS}}(a,c)$ such that WAS-LoRA selects e_2 whenever $\beta > \beta_{\mathrm{WAS}}(a,c)$. The corresponding rank-one update is

$$\widetilde{W}_{\mathrm{WAS}} = \mathrm{diag}(0, c).$$

Its finite endpoint error is $\mathcal{L}_{\mathrm{WAS}} = \mathbb{E}\left[\tanh\big(\beta + (1+a)x_1\big) - \tanh(\beta + x_1)\right]^2$. For $\beta > 1 + a$, the mean-value theorem gives

$$\mathcal{L}_{\mathrm{WAS}} \le \frac{a^2}{3} \mathrm{sech}^4\big(\beta - (1+a)\big) \longrightarrow 0. \tag{49}$$

Since $\delta_c > 0$, for sufficiently large but finite β, $\mathcal{L}_{\mathrm{WAS}} < \mathcal{L}_{\mathrm{SVD}}$. Thus the local-to-downstream reversal persists for a smooth activation: SVD retains the locally larger direction, while WAS-LoRA eventually selects the smaller direction because the former lies in the saturated region of the downstream nonlinearity. The inequalities are strict and the quantities involved depend continuously on the parameters, so the phenomenon persists under sufficiently small perturbations.

B.3 HIGHER-DIMENSIONAL NONLINEAR RANK COMPRESSION

The mismatch can affect an entire rank budget rather than a single direction. Consider $x \sim \mathrm{Unif}([-1,1]^{10}), \quad \mathbb{E}[xx^\top] = \frac{1}{3} I_{10}$, and the vector-valued endpoint

$$F_{\widetilde{W}}(x) = \tanh\big((I_{10} + \widetilde{W})x + b\big), \tag{50}$$

with componentwise $\tanh$, trained update

$$\Delta W = \mathrm{diag}(10, 10, 10, 10, 10, 1, 1, 1, 1, 1), \tag{51}$$

and $b = (15, 15, 15, 15, 15, 0, 0, 0, 0, 0)^\top$. We compress the adapter from rank ten to rank five.

Because the input covariance is isotropic, weight SVD and activation-aware local reconstruction both retain the first five directions:

$$\widetilde{W}_{\mathrm{SVD}} = \mathrm{diag}(10, 10, 10, 10, 10, 0, 0, 0, 0, 0). \tag{52}$$

Thus the entire local rank budget is assigned to the five directions with larger adapter magnitude.

Now use the LoRA factorization $A = I_{10}, \quad B = \Delta W, \quad \Sigma = \frac{1}{3} I_{10}$. Since the endpoint is coordinatewise, the whitened sensitivity matrix is diagonal:

$$M = \mathrm{diag}(\lambda_{\mathrm{sat}} I_5, \lambda_{\mathrm{act}} I_5), \tag{53}$$

where, for $X \sim \mathrm{Unif}[-1,1]$, $\lambda_{\mathrm{sat}} = \frac{100}{3}\mathbb{E}\,\mathrm{sech}^4(15 + 11X), \quad \lambda_{\mathrm{act}} = \frac{1}{3}\mathbb{E}\,\mathrm{sech}^4(2X)$. Because $15 + 11X \geq 4$,

$$\lambda_{\mathrm{sat}} \leq \frac{100}{3}\,\mathrm{sech}^4(4) < 6 \times 10^{-5}.$$

On the other hand,

$$\lambda_{\mathrm{act}} = \frac{1}{6}\left(\tanh 2 - \frac{1}{3}\tanh^3 2\right) \approx 0.1109.$$

Hence the five leading eigenvectors of M are exactly $e_6, \ldots, e_{10}$, and WAS-LoRA produces

$$\widetilde{W}_{\mathrm{WAS}} = \mathrm{diag}(0, 0, 0, 0, 0, 1, 1, 1, 1, 1). \tag{54}$$

This reversal is also reflected in the finite endpoint loss. The SVD solution discards the five unsaturated directions, giving

$$\mathcal{L}_{\mathrm{SVD}} = 5\mathbb{E}\left[\tanh(2X) - \tanh(X)\right]^2 \approx 0.291. \tag{55}$$

The WAS-LoRA solution discards only the saturated directions. Since $15 + 11X \geq 4$, the mean-value theorem gives

$$\mathcal{L}_{\mathrm{WAS}} \leq \frac{500}{3}\,\mathrm{sech}^4(4) < 3.0 \times 10^{-4}. \tag{56}$$

Thus, in this rank-10 to rank-5 example, activation-aware SVD allocates the entire rank budget to a five-dimensional subspace that is disjoint from the one selected by downstream sensitivity, with an endpoint-error gap of nearly three orders of magnitude.

The repeated diagonal values are used only to make the calculation transparent. Because both the sensitivity ordering and the finite-loss comparison are strict, the same reversal persists under sufficiently small perturbations of the weights and biases.

B.4 A TRANSFORMER REALIZATION OF THE LOCAL-TO-DOWNSTREAM REVERSAL

The local-to-downstream reversal can also occur inside a Transformer block, without relying on activation saturation. Consider two tokens, one attention head, and width two:

$$X = I_2, \qquad W_Q = W_K = I_2. \tag{57}$$

The genuine row-wise softmax attention matrix is

$$\mathcal{A} = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{2}}\right) = \begin{bmatrix} p & q \\ q & p \end{bmatrix}, \tag{58}$$

$$p = \frac{e^{1/\sqrt{2}}}{e^{1/\sqrt{2}} + 1}, \qquad q = 1 - p, \tag{59}$$

so both tokens mix nontrivially.

Place the trained rank-two LoRA update on the value projection and attenuate its first feature by the frozen output projection:

$$\Delta W_V = \mathrm{diag}(10, 1), \qquad W_O = \mathrm{diag}(\epsilon, 1), \qquad 0 < \epsilon < 0.1. \tag{60}$$

The adapter contribution at the block endpoint is therefore

$$\Delta Y = \mathcal{A} X \Delta W_V W_O = \mathcal{A}\,\mathrm{diag}(10\epsilon, 1). \tag{61}$$

Activation-aware SVD. Because $X = I_2$, weight SVD and activation-aware local reconstruction have the same exact rank-one optimum:

$$\widetilde{W}_{V,\mathrm{SVD}} = \mathrm{diag}(10, 0). \tag{62}$$

Thus local compression retains the first adapter direction and removes the second. The resulting squared block-output error is

$$E_{\mathrm{SVD}} = \|\mathcal{A}\,\mathrm{diag}(0, 1)\|_F^2 = p^2 + q^2. \tag{63}$$

WAS-LoRA at the block endpoint. Choose the equivalent LoRA factorization $A_V = I_2, \quad B_V = \mathrm{diag}(10, 1)$. Across the two token positions, the bottleneck activations are the two coordinate vectors, so under uniform token weighting $\Sigma = \frac{1}{2} I_2$. Let $u = \Sigma^{-1/2} s$ denote the whitened bottleneck coordinates. Since the query and key projections are frozen with respect to the value-adapter intervention, $\mathcal{A}$ is fixed when differentiating the chosen block endpoint.

For token position t, perturbing its whitened bottleneck changes the vectorized block output through $J_t^{\mathrm{w}} = \frac{\partial\, \mathrm{vec}(\Delta Y)}{\partial u_t}$. The corresponding Gram matrix is

$$(J_t^{\mathrm{w}})^\top J_t^{\mathrm{w}} = \frac{1}{2} \|\mathcal{A}_{:t}\|_2^2\, \mathrm{diag}(100\epsilon^2, 1) \tag{64}$$

$$= \frac{1}{2}(p^2 + q^2)\, \mathrm{diag}(100\epsilon^2, 1). \tag{65}$$

Both token positions give the same Gram matrix. Hence the whitened sensitivity matrix is, up to the common normalization convention for averaging token positions,

$$M_{\mathrm{block}} = \frac{1}{2}(p^2 + q^2)\, \mathrm{diag}(100\epsilon^2, 1). \tag{66}$$

The common positive scalar has no effect on its eigenspaces.

Since $\epsilon < 0.1$, $1 > 100\epsilon^2$, so the unique leading eigenvector of M_{block} is e_2. Thus WAS-LoRA retains the second bottleneck direction. Absorbing this direction into standard rank-one LoRA factors gives $B_{V,1} = B_V \Sigma^{1/2} e_2 = \frac{1}{\sqrt{2}} e_2, \quad A_{V,1} = e_2^\top \Sigma^{-1/2} A_V = \sqrt{2}\, e_2^\top$, and therefore

$$\widetilde{W}_{V,\mathrm{WAS}} = B_{V,1} A_{V,1} = \mathrm{diag}(0, 1). \tag{67}$$

The WAS-LoRA-compressed adapter removes only the first direction, whose effect is attenuated by W_O. Its squared block-output error is

$$E_{\mathrm{WAS}} = \|\mathcal{A}\,\mathrm{diag}(10\epsilon, 0)\|_F^2 = 100\epsilon^2 (p^2 + q^2). \tag{68}$$

Consequently,

$$\frac{E_{\mathrm{SVD}}}{E_{\mathrm{WAS}}} = \frac{1}{100\epsilon^2} \longrightarrow \infty \qquad \text{as } \epsilon \to 0. \tag{69}$$

Thus, in this genuine softmax-attention block, activation-aware SVD and WAS-LoRA select opposite rank-one subspaces:

$$\widetilde{W}_{V,\mathrm{SVD}} = \mathrm{diag}(10, 0), \qquad \widetilde{W}_{V,\mathrm{WAS}} = \mathrm{diag}(0, 1).$$

The reversal does not arise from a nonlinear activation or a closed gate; it is induced solely by the frozen downstream output projection, which changes the geometry relevant to the chosen endpoint. Residual connections and the frozen base value projection are shared by teacher and compressed model and therefore cancel in their difference.

C Analysis of subspace selection

C.1 Deletion tests on matched candidate directions

We assess candidate scores by comparing their rankings with the observed effects of deleting individual rank-one LoRA components. We use three rank-16 SST-2 teachers independently trained with random seeds 0,1,2. For each teacher, compact SVD of the 24 query/value updates yields $24 \times 16 = 384$ rank-one SVD candidates. Local singular energy and Candidate-WAS rank these identical candidates; neither may rotate them. Candidate-WAS measures the average squared first-order effect on raw logits on a calibration split disjoint from the audit data. For each candidate, we then delete only that component from the otherwise intact network and measure the exact change in raw logits on 128 held-out examples. The deletion effect is never used to construct the score.

Across three teachers trained with random seeds 0, 1, and 2, the Spearman correlations with held-out deletion effects are respectively 0.8624, 0.8573, and 0.8633 for local singular energy, and 0.9586, 0.9552, and 0.9574 for Candidate-WAS (Figure 5).

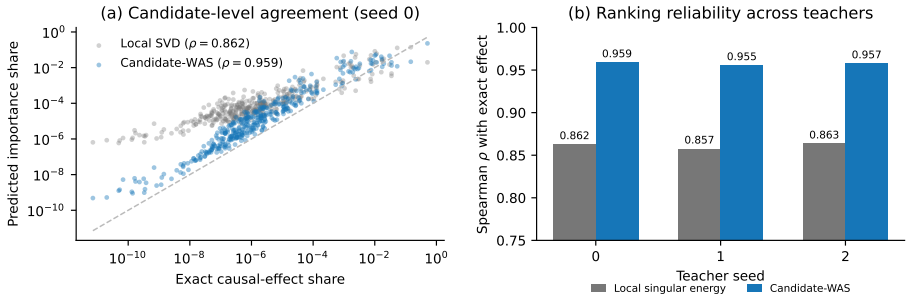


Figure 5: Candidate scores versus held-out deletion effects. Both scores use the same 384 physical SVD candidates. Candidate-WAS tracks exact finite single-deletion effects more closely, while unrestricted WAS-LoRA may additionally rotate the retained subspace.

At a total retained-rank budget of $K = 96$, each scoring rule selects 96 components from the shared candidate pool. The selected sets overlap by 78.5% on average across the three teachers, where overlap is defined as the intersection size divided by 96. Thus, Local SVD and Candidate-WAS retain many of the same components, but their selections are not identical.

This fixed-candidate experiment tests scoring quality, not the end-to-end performance of full WAS-LoRA or the benefit of constructing new directions.

C.2 Output-scored Act-SVD

The control experiment uses the same teachers, calibration inputs, downstream raw-logit Gram matrices, total ranks, and rank-one floor as WAS-LoRA. For each module, it constructs the ridge-free support-whitened local basis U_0 and scores its fixed directions by the diagonal of $U_0^\top M_{\text{logits}} U_0$. It reserves one direction per module, then fills the budget by global score ordering. It may skip locally higher-energy directions but cannot form new basis vectors. WAS-LoRA instead constructs unrestricted leading active directions.

Across the five tasks, eight independently trained teachers per task, and five retained-rank budgets, the geometric-mean ratios of centered relative logit error are $0.8302\ [0.7883,\ 0.8690]$ for Output-scored Act-SVD relative to Act-SVD and $0.8422\ [0.8338,\ 0.8506]$ for WAS-LoRA relative to Output-scored Act-SVD. Brackets indicate paired-seed 95% descriptive intervals.

Table 2 reports the aggregate comparison together with representative per-task results at average retained rank $\bar{r} = 4$. At this budget, the same ordering Act-SVD $\rightarrow$ Output-scored Act-SVD $\rightarrow$ WAS-LoRA is observed on all five tasks.

The results separate two effects. Giving the fixed activation-aware SVD basis access to downstream information already improves teacher-function preservation: Output-scored Act-SVD reduces the

Table 2: Output-scored Act-SVD control. Top: geometric-mean ratios of centered relative logit error across five tasks, eight independently trained teachers per task, and five retained-rank budgets; brackets denote paired-seed 95% descriptive intervals. Bottom: centered relative logit error at average retained rank $\bar{r} = 4$, reported as mean ± sample standard deviation across eight teachers. Lower is better.

Aggregate comparison across all five budgets		
Comparison	**Geometric-mean ratio**	**95% interval**
Output-scored Act-SVD / Act-SVD	0.8302	[0.7883, 0.8690]
WAS-LoRA / Output-scored Act-SVD	0.8422	[0.8338, 0.8506]

Centered relative logit error at $\bar{r} = 4$			
Task	**Act-SVD**	**Output-scored Act-SVD**	**WAS-LoRA**
SST-2	0.0523 ± 0.0078	0.0475 ± 0.0069	0.0427 ± 0.0069
QNLI	0.0787 ± 0.021	0.0740 ± 0.023	0.0684 ± 0.022
MRPC	0.0292 ± 0.014	0.0237 ± 0.012	0.0179 ± 0.0092
RTE	0.0694 ± 0.022	0.0487 ± 0.027	0.0409 ± 0.025
MNLI	0.158 ± 0.064	0.146 ± 0.057	0.139 ± 0.056

aggregate centered-logit error relative to Act-SVD. Allowing the downstream geometry to construct an unrestricted active subspace yields a further reduction, as reflected by the WAS-LoRA/Output-scored Act-SVD ratio and by the consistent per-task ordering in Table 2.

This comparison is not, however, a pure fixed-allocation rotation ablation. Even under matched total rank budgets, per-module ranks can differ. Moreover, WAS-LoRA uses a small ridge term, whereas the local-basis controls use ridge-free whitening on the data-supported subspace. The additional gain over Output-scored Act-SVD should therefore be interpreted as evidence for the value of unrestricted downstream-aware subspace construction under the full WAS-LoRA procedure, rather than as being attributable solely to basis rotation. Complete per-task results across all five retained-rank budgets are reported in Appendix D.1.2 for centered relative logit error and Appendix D.1.3 for teacher-to-compressed KL.

C.3 Calibration-set size sensitivity

The main encoder experiments use fixed, task-specific calibration sets. We assess their size sensitivity at two complementary levels. First, a four-task sweep measures how the selected compression subspaces and module-wise rank allocations change as the number of calibration examples grows. Second, an end-to-end MNLI sweep measures the accuracy and teacher-output fidelity of the resulting compressed models. No recovery training is used in either study.

Geometry across tasks and budgets. We use SST-2, QNLI, MRPC, and RTE, whose fixed calibration-pool sizes are 2,048, 2,048, 512, and 384, respectively. For each task, we evaluate three frozen teachers, five total-rank budgets $K \in \{48, 72, 96, 144, 192\}$, and ten fixed nested subset draws at every feasible non-full size. The subsets follow the same class-proportional construction as the paper calibration sets.

We compare each subset compression with its corresponding full-pool WAS-LoRA compression using two complementary metrics. Let $Q_{m,s}^{(N)}$ and $Q_{m,s}^{(N_{\text{full}})}$ be orthonormal bases for side $s \in \{L, R\}$ of the physical update in module m, with dimensions $d_{m,s}^{(N)}$ and $d_{m,s}^{(N_{\text{full}})}$. The physical-subspace error is

$$e_{\text{sub}}(N) = 1 - \frac{\sum_{m,s} 2\|(Q_{m,s}^{(N)})^\top Q_{m,s}^{(N_{\text{full}})}\|_F^2}{\sum_{m,s}(d_{m,s}^{(N)} + d_{m,s}^{(N_{\text{full}})})}. \tag{70}$$

It tracks changes in the represented left and right subspaces while also penalizing dimension changes. If $r_m^{(N)}$ is the rank assigned to module m, the normalized rank-transport distance is

$$e_{\text{rank}}(N) = \frac{1}{2K} \sum_m \left| r_m^{(N)} - r_m^{(N_{\text{full}})} \right|. \tag{71}$$

Both quantities are zero at the full-pool endpoint by construction, so that endpoint is omitted from the curves. Lines are means over three teachers and ten draws; bands are descriptive 10th–90th percentile ranges.

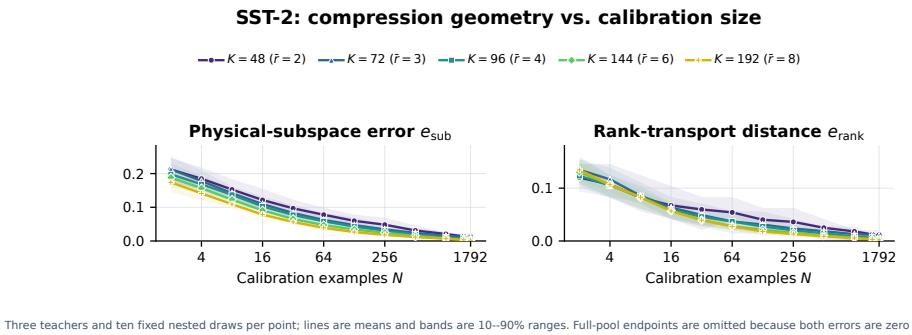


Figure 6: Calibration-size sensitivity of the WAS-LoRA compression geometry on SST-2. The panels show physical-subspace error and normalized rank-transport distance; colors denote the five total-rank budgets.

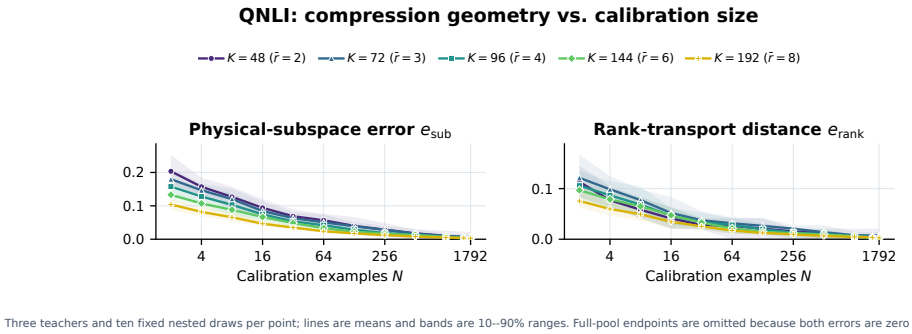


Figure 7: Calibration-size sensitivity of the WAS-LoRA compression geometry on QNLI, with the conventions of Figure 6.

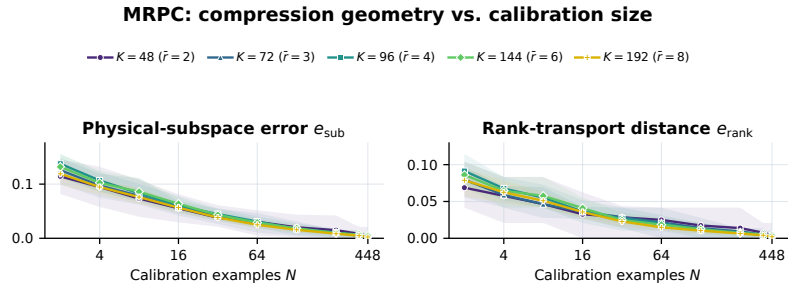


Figure 8: Calibration-size sensitivity of the WAS-LoRA compression geometry on MRPC, with the conventions of Figure 6.

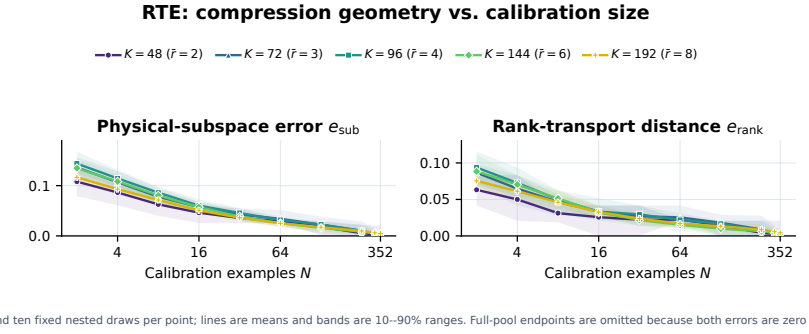


Figure 9: Calibration-size sensitivity of the WAS-LoRA compression geometry on RTE, with the conventions of Figure 6.

Across all four tasks, the curves exhibit early and gradual stabilization (Figures 6 to 9). At the smallest common size, $N = 16$, mean physical-subspace error is at most 0.122 and mean rank-transport distance at most 0.067 across all 20 task–budget cells. At $N = 64$, the corresponding means lie in $[0.024, 0.078]$ and $[0.015, 0.054]$. Their medians decrease from $0.062/0.039$ at $N = 16$ to $0.032/0.024$ at $N = 64$, and then to $0.014/0.011$ at $N = 256$ (subspace/rank error). Larger budgets generally reach the low-error regime earlier, while even the most restrictive budget $K = 48$ stabilizes steadily as more examples are added.

End-to-end MNLI size sweep. We vary the MNLI calibration size over $N \in \{32, 64, 128, 256, 2048\}$, where $N = 2048$ is the finite full-pool reference. The fixed MNLI partition contains 388,606 teacher-training examples, 2,048 checkpoint-selection examples, and a disjoint calibration pool of 2,048 training examples; the official matched and mismatched validation splits are used only for evaluation. The pool contains 682, 683, and 683 examples from the three classes.

For each of five prespecified draws (seeds 0–4), subsets are sampled without replacement to match the class proportions of the fixed pool and satisfy $S_{32} \subset S_{64} \subset S_{128} \subset S_{256} \subset S_{2048}$. Labels are used only to construct these subsets and are not passed to the WAS-LoRA statistic collector or rank allocator. We evaluate three frozen teachers (seeds 0–2) at the primary budget $K = 192$ and the tighter budget $K = 48$; both budgets reuse exactly the same subset statistics. The task head and original LoRA scaling are unchanged. Metrics are computed separately on the matched and mismatched validation splits and then averaged with equal split weight. We average the five draws within each teacher and then the three teacher means. The full-pool reference is one unique model per teacher rather than five repeated runs.

Primary budget ($K = 192$). The paired curves in Figure 4 remain close to the full-pool reference throughout the measured range. As detailed in Table 3, every non-full mean accuracy is within 0.045 percentage points of the reference. The largest absolute mean changes in centered teacher-logit error and KL are 7.4×10^{-4} and 1.2×10^{-4}, respectively. Thus, even $N = 32$, which uses only 1.56% of the full pool, preserves both task accuracy and teacher-output fidelity in this setting.

Tighter budget ($K = 48$). Task accuracy remains close to the full-pool reference under substantially stronger compression (Table 4): its mean difference is -0.143 percentage points at $N = 32$ and at most 0.077 points in magnitude for $N \geq 64$. Across the non-full sizes, the maximum centered-error and KL gaps are only 0.00874 and 0.00475, respectively. Thus the same qualitative stability persists under the substantially tighter compression budget.

The direct centered-logit distance between each subset-compressed model and its corresponding full-pool compressed model decreases as the calibration set grows. Its mean values for $N = 32, 64, 128, 256$ are $0.0215, 0.0177, 0.0156, 0.0127$ at $K = 192$ and $0.0810, 0.0664, 0.0552, 0.0474$ at $K = 48$. This distance is normalized by the centered teacher-logit norm and equals zero at $N = 2048$ by construction.

Competitive accuracy–fidelity frontier at $N = 64$. The small-calibration models also remain competitive with the seven paper controls (Figure 10). At $K = 48$, WAS-LoRA has the highest mean accuracy, exceeding the strongest control by 0.035 percentage points, while remaining close to the strongest centered-logit fidelity. At $K = 192$, it reduces centered teacher-logit error by 14.8% relative to the next-lowest mean and KL by 13.6%, while its accuracy is within 0.040 percentage points of the strongest-accuracy control. Thus the fidelity advantage motivating WAS-LoRA survives a 32-fold calibration-data reduction without a material accuracy cost.

Together, the cross-task geometry sweep and the end-to-end MNLI evaluation show that WAS-LoRA is robust to calibration-set size in the evaluated setting: its induced geometry stabilizes with relatively few examples, while aggressive subsampling preserves both teacher-output fidelity and task accuracy.

Table 3: MNLI calibration-size sensitivity for WAS-LoRA at $K = 192$. Absolute metrics are mean $\pm$ sample standard deviation across the three teacher means after averaging five fixed draws within teacher; the full reference has one unique model per teacher. Delta columns are paired small-minus-full means. Matched and mismatched validation splits receive equal weight.

N	Accuracy (%) ↑	Δ Acc. (pp)	e_{ctr} ↓	Δe_{ctr}	KL ↓	Δ KL
32	86.583 ± 0.112	$+0.044$	0.03755 ± 0.03139	-0.00020	0.002208 ± 0.003096	$+0.000116$
64	86.579 ± 0.101	$+0.040$	0.03761 ± 0.03163	-0.00015	0.002153 ± 0.003008	$+0.000061$
128	86.581 ± 0.105	$+0.042$	0.03701 ± 0.03097	-0.00074	0.002094 ± 0.002919	$+0.000002$
256	86.564 ± 0.106	$+0.025$	0.03748 ± 0.03142	-0.00027	0.002115 ± 0.002942	$+0.000023$
2,048	86.539 ± 0.074	0.000	0.03775 ± 0.03202	0.00000	0.002092 ± 0.002918	0.000000

Table 4: MNLI calibration-size sensitivity at the tighter budget $K = 48$, using the conventions of Table 3.

N	Accuracy (%) ↑	Δ Acc. (pp)	e_{ctr} ↓	Δe_{ctr}	KL ↓	Δ KL
32	84.717 ± 0.632	-0.143	0.30915 ± 0.08839	$+0.00862$	0.06942 ± 0.03627	$+0.00475$
64	84.782 ± 0.573	-0.077	0.30927 ± 0.08963	$+0.00874$	0.06763 ± 0.03485	$+0.00296$
128	84.823 ± 0.574	-0.037	0.30549 ± 0.08949	$+0.00495$	0.06585 ± 0.03418	$+0.00119$
256	84.829 ± 0.499	-0.031	0.30875 ± 0.08275	$+0.00821$	0.06645 ± 0.03256	$+0.00178$
2,048	84.860 ± 0.606	0.000	0.30053 ± 0.08043	0.00000	0.06467 ± 0.03266	0.00000

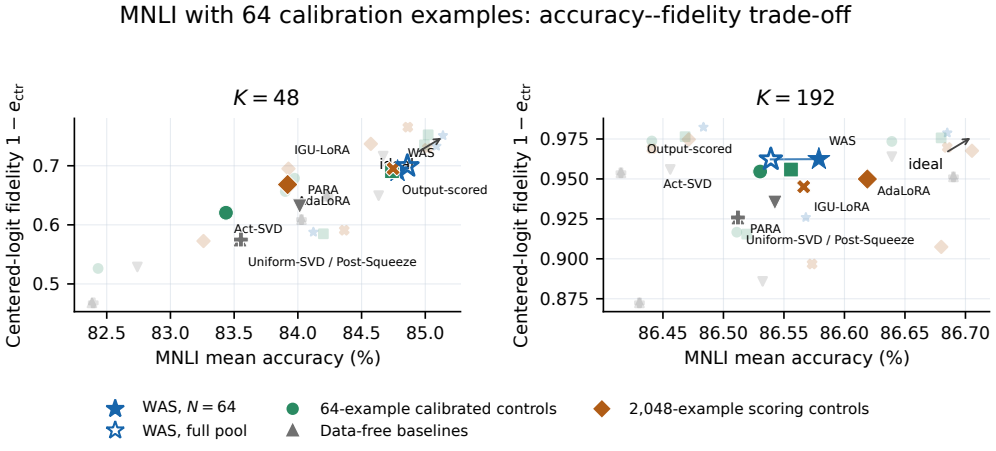


Figure 10: Accuracy–fidelity trade-off on MNLI with $N = 64$. The vertical coordinate is centered teacher-logit fidelity $1 - e_{\mathrm{ctr}}$, so the upper right is better. Large markers are equal-teacher means and faint markers are the three teacher means. Hollow stars show full-pool WAS-LoRA and filled stars show 64-example WAS-LoRA. Data-free baselines and labeled scoring controls retain their original calibration protocols.

D ADDITIONAL RESULTS

D.1 ABSOLUTE METRICS

All tables use the same 40 teachers and all five retained-rank budgets. Tables 5 to 7 report arithmetic means and sample standard deviations across eight seeds for each task. MRPC uses raw accuracy; its official score averages accuracy and F1. MNLI averages matched and mismatched results within each teacher cell before aggregation. Official scores are retained in `data/per_task_metrics.csv` and `data/teacher_cells.csv`. In the absolute tables, Output-scored denotes Output-scored Act-SVD; Act-SVD denotes the exact local $H = 0$ control.

D.1.1 ACCURACY

Table 5: Raw classification accuracy (percent). Mean ± sample standard deviation over eight teachers per task.

Method	$r = 2$	$r = 3$	$r = 4$	$r = 6$	$r = 8$
SST-2					
Uniform-SVD	93.22 ± 0.23	93.49 ± 0.51	93.59 ± 0.41	93.74 ± 0.51	93.76 ± 0.47
PARA	93.05 ± 0.35	93.43 ± 0.37	93.64 ± 0.48	93.81 ± 0.40	93.78 ± 0.52
Post-Squeeze	93.22 ± 0.23	93.49 ± 0.51	93.59 ± 0.41	93.74 ± 0.51	93.76 ± 0.47
Act-SVD	93.33 ± 0.22	93.53 ± 0.41	93.58 ± 0.38	93.76 ± 0.55	93.78 ± 0.59
Output-scored	93.29 ± 0.25	93.56 ± 0.44	93.64 ± 0.49	93.71 ± 0.52	93.76 ± 0.55
WAS	93.32 ± 0.22	93.61 ± 0.24	93.71 ± 0.52	93.71 ± 0.54	93.72 ± 0.56
AdaLoRA-score	93.31 ± 0.48	93.45 ± 0.46	93.62 ± 0.42	93.68 ± 0.43	93.82 ± 0.46
IGU-LoRA-score	93.08 ± 0.35	93.48 ± 0.36	93.65 ± 0.43	93.79 ± 0.47	93.82 ± 0.46
QNLI					
Uniform-SVD	91.00 ± 0.60	91.74 ± 0.25	92.03 ± 0.24	92.30 ± 0.22	92.29 ± 0.24
PARA	91.29 ± 0.48	91.87 ± 0.24	92.20 ± 0.23	92.27 ± 0.27	92.32 ± 0.25
Post-Squeeze	91.00 ± 0.60	91.74 ± 0.26	92.03 ± 0.24	92.30 ± 0.22	92.29 ± 0.24
Act-SVD	91.49 ± 0.41	91.90 ± 0.23	92.23 ± 0.25	92.36 ± 0.25	92.34 ± 0.26
Output-scored	91.70 ± 0.28	92.09 ± 0.21	92.26 ± 0.20	92.32 ± 0.24	92.36 ± 0.26
WAS	91.76 ± 0.28	92.15 ± 0.21	92.24 ± 0.25	92.31 ± 0.27	92.34 ± 0.27
AdaLoRA-score	91.67 ± 0.36	92.05 ± 0.25	92.24 ± 0.24	92.34 ± 0.26	92.38 ± 0.26
IGU-LoRA-score	91.66 ± 0.29	92.06 ± 0.26	92.20 ± 0.25	92.30 ± 0.28	92.35 ± 0.25
MRPC					
Uniform-SVD	87.38 ± 1.12	87.68 ± 1.06	87.87 ± 1.13	87.90 ± 1.16	88.05 ± 1.09
PARA	87.56 ± 1.36	87.44 ± 1.24	87.71 ± 1.43	87.81 ± 1.23	87.93 ± 1.18
Post-Squeeze	87.38 ± 1.12	87.68 ± 1.06	87.87 ± 1.13	87.90 ± 1.16	88.05 ± 1.09
Act-SVD	87.65 ± 0.91	87.81 ± 1.14	87.75 ± 1.13	87.96 ± 1.07	87.93 ± 1.11
Output-scored	87.84 ± 1.24	87.99 ± 1.11	87.81 ± 1.22	87.99 ± 1.17	87.96 ± 1.15
WAS	87.78 ± 1.17	88.02 ± 1.23	87.99 ± 1.31	88.05 ± 1.25	87.99 ± 1.24
AdaLoRA-score	87.71 ± 0.95	87.90 ± 1.27	87.87 ± 1.21	87.99 ± 1.06	88.02 ± 1.12
IGU-LoRA-score	87.68 ± 0.92	87.87 ± 1.10	87.87 ± 1.12	87.96 ± 1.10	88.08 ± 1.10
RTE					
Uniform-SVD	71.89 ± 2.70	72.47 ± 1.28	72.79 ± 1.77	72.83 ± 2.39	72.88 ± 2.29
PARA	71.66 ± 2.46	72.02 ± 1.46	72.25 ± 1.69	72.79 ± 1.79	72.92 ± 2.08
Post-Squeeze	71.89 ± 2.70	72.47 ± 1.28	72.79 ± 1.77	72.83 ± 2.39	72.88 ± 2.29
Act-SVD	72.65 ± 1.79	72.61 ± 1.73	73.01 ± 1.72	72.74 ± 2.25	73.10 ± 2.11
Output-scored	72.70 ± 1.74	72.47 ± 2.71	72.83 ± 2.34	73.06 ± 2.10	73.06 ± 2.03
WAS	73.01 ± 1.53	72.70 ± 2.10	73.06 ± 1.94	72.97 ± 2.06	72.88 ± 2.15
AdaLoRA-score	71.98 ± 1.53	72.16 ± 2.15	72.92 ± 2.05	73.01 ± 2.16	72.97 ± 2.06
IGU-LoRA-score	72.11 ± 1.56	72.11 ± 2.09	72.97 ± 2.03	72.88 ± 2.29	72.88 ± 2.19
MNLI					
Uniform-SVD	83.32 ± 0.89	84.86 ± 0.54	85.55 ± 0.35	86.22 ± 0.15	86.40 ± 0.17
PARA	83.41 ± 1.41	85.19 ± 0.52	85.84 ± 0.37	86.36 ± 0.21	86.50 ± 0.09
Post-Squeeze	83.32 ± 0.89	84.86 ± 0.54	85.55 ± 0.34	86.22 ± 0.15	86.40 ± 0.17
Act-SVD	83.05 ± 1.45	85.12 ± 0.51	85.87 ± 0.31	86.36 ± 0.16	86.48 ± 0.08
Output-scored	84.50 ± 0.45	85.48 ± 0.30	85.98 ± 0.24	86.43 ± 0.16	86.54 ± 0.07
WAS	84.62 ± 0.50	85.60 ± 0.25	86.10 ± 0.28	86.42 ± 0.15	86.55 ± 0.07
AdaLoRA-score	84.02 ± 0.56	85.52 ± 0.28	86.09 ± 0.18	86.47 ± 0.14	86.61 ± 0.11
IGU-LoRA-score	84.46 ± 0.64	85.57 ± 0.38	86.08 ± 0.20	86.45 ± 0.10	86.54 ± 0.11

D.1.2 TEACHER-OUTPUT ERROR

Table 6: Centered relative logit error. Mean ± sample standard deviation over eight teachers per task. Uniform = Uniform-SVD; Post-Sq. = Post-Squeeze; Output-sc. = Output-scored Act-SVD; Ada/IGU-score = AdaLoRA/IGU-LoRA-score. Leading decimal zeros are omitted.

Method	$r = 2$	$r = 3$	$r = 4$	$r = 6$	$r = 8$
SST-2					
Uniform	.137±.013	.08±.0071	.0576±.0064	.0336±.0053	.0212±.0032
PARA	.158±.023	.0932±.0092	.068±.0068	.0352±.0067	.0201±.0017
Post-Sq.	.137±.013	.08±.0071	.0576±.0064	.0336±.0053	.0212±.0032
Act-SVD	.137±.025	.0779±.0091	.0523±.0078	.0267±.0043	.0154±.0018
Output-sc.	.113±.011	.0725±.0076	.0475±.0069	.0246±.0027	.0132±.0011
WAS	.106±.0094	.0668±.0067	.0427±.0069	.0207±.0022	.0109±.0011
Ada-score	.124±.011	.0771±.0098	.0531±.007	.0278±.0035	.0157±.0021
IGU-score	.137±.018	.0846±.0094	.0592±.0069	.0302±.0039	.0158±.0017
QNLI					
Uniform	.277±.062	.162±.045	.112±.033	.0592±.018	.0356±.0088
PARA	.266±.072	.161±.044	.105±.03	.053±.017	.029±.0077
Post-Sq.	.277±.062	.162±.045	.112±.033	.0592±.018	.0356±.0088
Act-SVD	.224±.063	.129±.033	.0787±.021	.039±.011	.0213±.0056
Output-sc.	.208±.056	.118±.035	.074±.023	.0338±.0095	.018±.0048
WAS	.189±.046	.11±.032	.0684±.022	.0294±.0086	.0145±.004
Ada-score	.217±.057	.12±.037	.0774±.023	.0395±.011	.0214±.0064
IGU-score	.229±.06	.136±.043	.092±.029	.0408±.013	.0217±.0062
MRPC					
Uniform	.111±.028	.0546±.014	.0393±.013	.0208±.0087	.0116±.0047
PARA	.129±.047	.0736±.021	.0487±.0079	.0254±.0059	.0157±.0039
Post-Sq.	.111±.028	.0546±.014	.0393±.013	.0208±.0087	.0116±.0047
Act-SVD	.0844±.046	.044±.023	.0292±.014	.0163±.0072	.00954±.0044
Output-sc.	.0872±.052	.0388±.021	.0237±.012	.0129±.007	.00741±.0033
WAS	.0756±.047	.0296±.016	.0179±.0092	.00872±.0042	.00473±.0023
Ada-score	.0691±.034	.0361±.019	.0259±.013	.0146±.0083	.00758±.004
IGU-score	.0648±.034	.0384±.021	.027±.016	.0135±.0073	.00799±.0046
RTE					
Uniform	.304±.11	.146±.048	.1±.041	.0489±.024	.029±.016
PARA	.336±.11	.206±.062	.115±.039	.0596±.018	.0284±.011
Post-Sq.	.304±.11	.146±.048	.1±.041	.0489±.024	.029±.016
Act-SVD	.188±.071	.103±.032	.0694±.022	.0357±.016	.0225±.01
Output-sc.	.126±.063	.0752±.044	.0487±.027	.027±.017	.0163±.01
WAS	.11±.059	.0642±.041	.0409±.025	.0219±.014	.0116±.0075
Ada-score	.144±.054	.0805±.041	.0511±.028	.0284±.017	.0165±.0087
IGU-score	.154±.06	.085±.043	.0554±.025	.0289±.017	.0175±.011
MNLI					
Uniform	.45±.093	.285±.1	.215±.086	.13±.054	.0844±.041
PARA	.408±.11	.261±.11	.189±.085	.118±.059	.0727±.041
Post-Sq.	.45±.093	.285±.1	.215±.086	.13±.054	.0844±.041
Act-SVD	.398±.11	.231±.085	.158±.064	.0865±.045	.0519±.029
Output-sc.	.339±.097	.212±.073	.146±.057	.0813±.041	.0491±.028
WAS	.319±.086	.201±.062	.139±.056	.0743±.04	.0438±.027
Ada-score	.349±.092	.223±.082	.157±.07	.09±.045	.0543±.031
IGU-score	.327±.094	.224±.09	.167±.081	.0971±.051	.0634±.038

D.1.3 KL DIVERGENCE

Table 7: Teacher-to-compressed KL divergence. Mean ± sample standard deviation over eight teachers per task. Uniform = Uniform-SVD; Post-Sq. = Post-Squeeze; Output-sc. = Output-scored Act-SVD; Ada/IGU-score = AdaLoRA/IGU-LoRA-score. Leading decimal zeros are omitted. $a\mathrm{e}{-b}$ denotes $a \times 10^{-b}$.

Method	$r = 2$	$r = 3$	$r = 4$	$r = 6$	$r = 8$
SST-2					
Uniform	.0164±.0035	.00724±.0024	.00373±.0011	.00152±.0006	.000644±.00023
PARA	.0182±.0063	.00751±.0024	.00365±.0012	.00116±.00046	.000407±.00011
Post-Sq.	.0164±.0035	.00724±.0024	.00373±.0011	.00152±.0006	.000644±.00023
Act-SVD	.0188±.0077	.00667±.0017	.00321±.0013	.000942±.00038	.000318±.00012
Output-sc.	.0121±.0035	.00582±.0021	.00268±.0011	.000766±.00024	.000227±6.8e−5
WAS	.0115±.003	.0051±.0015	.00231±.00093	.000562±.00014	.000161±4.6e−5
Ada-score	.0126±.0031	.00551±.0014	.00286±.00087	.000843±.0003	.000293±.00012
IGU-score	.0139±.004	.00577±.0015	.00315±.0012	.000897±.00031	.000271±7.5e−5
QNLI					
Uniform	.043±.012	.0179±.0073	.0101±.0049	.00308±.0014	.00122±.00047
PARA	.0359±.012	.0149±.0053	.00657±.0025	.00221±.0013	.000643±.00026
Post-Sq.	.043±.012	.0179±.0073	.0101±.0049	.00308±.0014	.00122±.00047
Act-SVD	.0306±.011	.0121±.0045	.00508±.002	.00148±.00064	.000444±.00018
Output-sc.	.0233±.0081	.00839±.0031	.00396±.0017	.000998±.00049	.00029±.00011
WAS	.02±.0066	.00756±.0029	.00339±.0015	.00078±.00034	.000186±7.2e−5
Ada-score	.0233±.0082	.00901±.0035	.00431±.0018	.00113±.00042	.000316±.00013
IGU-score	.0253±.0087	.0103±.0043	.00477±.0021	.00122±.00053	.000325±.00012
MRPC					
Uniform	.0207±.013	.0049±.0035	.00297±.0024	.000898±.00086	.000287±.00027
PARA	.0206±.014	.00691±.0037	.0031±.0015	.00085±.00035	.00039±.00025
Post-Sq.	.0207±.013	.0049±.0035	.00297±.0024	.000898±.00086	.000287±.00027
Act-SVD	.0112±.01	.00327±.0031	.0016±.0014	.000439±.00032	.000167±.00014
Output-sc.	.00728±.0058	.00225±.002	.00109±.0012	.000308±.00037	8.64e−5±7.4e−5
WAS	.00597±.0053	.00146±.0014	.000543±.00048	.000137±.00013	3.62e−5±3e−5
Ada-score	.00747±.0065	.00236±.0023	.00103±.00099	.00035±.00038	.000107±.00012
IGU-score	.00659±.0059	.00238±.0024	.00127±.0015	.000261±.00024	.000102±.00012
RTE					
Uniform	.149±.12	.0359±.037	.0193±.021	.00606±.01	.00276±.0048
PARA	.143±.11	.0562±.044	.0186±.018	.00591±.0075	.00157±.0023
Post-Sq.	.149±.12	.0359±.037	.0193±.021	.00606±.01	.00276±.0048
Act-SVD	.0669±.083	.0175±.02	.00823±.0095	.00283±.0045	.00121±.0022
Output-sc.	.0255±.033	.0132±.024	.00604±.011	.00246±.0048	.00111±.0024
WAS	.0212±.034	.00896±.016	.00317±.0048	.00139±.0027	.000642±.0015
Ada-score	.0356±.05	.0181±.036	.0064±.011	.0028±.0059	.000924±.0018
IGU-score	.0428±.059	.0178±.034	.00621±.01	.003±.0063	.00122±.0025
MNLI					
Uniform	.133±.055	.0601±.033	.0375±.023	.0168±.012	.00826±.0065
PARA	.118±.062	.0531±.035	.0306±.022	.0131±.011	.00532±.0047
Post-Sq.	.133±.055	.0601±.033	.0375±.023	.0168±.012	.00826±.0065
Act-SVD	.121±.059	.0502±.029	.0269±.018	.00985±.0083	.00392±.0035
Output-sc.	.083±.037	.0402±.022	.0216±.015	.00833±.0069	.00309±.0027
WAS	.0761±.033	.0367±.019	.0199±.014	.00728±.0062	.00269±.0025
Ada-score	.0923±.038	.0408±.024	.0227±.016	.00866±.007	.0035±.003
IGU-score	.0773±.037	.0397±.025	.0233±.017	.00927±.0079	.00417±.0037

D.2 Five-task mean accuracy

Table 8 gives equal-task mean raw accuracy and descriptive 95% intervals for all methods and budgets.

Table 8: Five-task mean raw accuracy (percent). Each cell gives the mean above its descriptive 95% interval. The two panels cover all five average retained ranks. Output-scored denotes Output-scored Act-SVD.

Method	$r = 2$	$r = 3$	$r = 4$
Uniform-SVD	85.3595 [85.0255, 85.6511]	86.0490 [85.8701, 86.2487]	86.3666 [86.1491, 86.6007]
PARA	85.3926 [85.0565, 85.7252]	85.9916 [85.8523, 86.1203]	86.3270 [86.0844, 86.6069]
Post-Squeeze	85.3601 [85.0268, 85.6515]	86.0492 [85.8695, 86.2498]	86.3666 [86.1490, 86.6015]
Act-SVD	85.6374 [85.3545, 85.9231]	86.1940 [85.9177, 86.4473]	86.4891 [86.2452, 86.7340]
Output-scored	86.0037 [85.6820, 86.3046]	86.3207 [85.8907, 86.7344]	86.5048 [86.1472, 86.8096]
WAS	86.0974 [85.8214, 86.3578]	86.4153 [86.0819, 86.7090]	86.6185 [86.3117, 86.8950]
AdaLoRA-score	85.7377 [85.4936, 85.9828]	86.2155 [85.8573, 86.5620]	86.5498 [86.2222, 86.8324]
IGU-LoRA-score	85.7984 [85.5847, 86.0020]	86.2188 [85.8836, 86.4861]	86.5527 [86.2174, 86.8583]

Method	$r = 6$	$r = 8$
Uniform-SVD	86.5978 [86.2469, 86.9093]	86.6769 [86.3340, 86.9805]
PARA	86.6064 [86.3986, 86.8300]	86.6898 [86.3874, 86.9774]
Post-Squeeze	86.5978 [86.2469, 86.9093]	86.6769 [86.3340, 86.9805]
Act-SVD	86.6365 [86.3227, 86.9249]	86.7273 [86.4047, 87.0270]
Output-scored	86.7016 [86.3575, 87.0105]	86.7374 [86.3995, 87.0730]
WAS	86.6931 [86.3543, 87.0071]	86.6960 [86.3442, 87.0468]
AdaLoRA-score	86.6977 [86.3515, 86.9943]	86.7595 [86.4189, 87.0830]
IGU-LoRA-score	86.6759 [86.3212, 86.9976]	86.7338 [86.3778, 87.0653]

D.3 Paired WAS/comparator summaries

Each rank comparison uses five fixed tasks and eight paired seed blocks. For fidelity, we average $\log(e_{\mathrm{WAS}}/e_b)$ equally across tasks within seed, then across seeds, and exponentiate. All-budget summaries also weight ranks equally before the seed average. KL uses the same rule. Accuracy differences are arithmetic task means in percentage points.

The descriptive 95% percentile bootstrap resamples eight seed IDs with replacement, coupling each seed ID across all tasks, methods, and budgets (10,000 draws; RNG seed 20260914). Tasks are not resampled. Win/tie/loss counts use paired teacher cells and tolerance 10^{-12}; they are descriptive, not independent trials. No confirmatory p-values or multiplicity-adjusted claims are made. Tables 9 to 11 retain all seven comparators and all five budgets. Below-one ratios and positive accuracy differences favor WAS. Complete official-score differences remain in `data/paired_comparisons.csv`; mean-accuracy intervals are in `data/accuracy_curves.csv`.

Table 9: Paired WAS/comparator centered-error ratio. Descriptive 95% intervals and W/T/L across 40 paired teacher cells per budget. Below one favors WAS.

Comparator	Rank	Estimate	95% interval	W/T/L
Uniform-SVD	2	0.5839	[0.5119, 0.6706]	39/0/1
	3	0.5932	[0.5465, 0.6477]	40/0/0
	4	0.5373	[0.5073, 0.5707]	40/0/0
	6	0.4876	[0.4606, 0.5133]	40/0/0
	8	0.4302	[0.4091, 0.4545]	40/0/0
PARA	2	0.5649	[0.5029, 0.6453]	39/0/1
	3	0.5172	[0.4680, 0.5782]	39/0/1
	4	0.4954	[0.4583, 0.5422]	40/0/0
	6	0.4592	[0.4155, 0.5070]	40/0/0
	8	0.4340	[0.4020, 0.4749]	40/0/0
Post-Squeeze	2	0.5839	[0.5119, 0.6706]	39/0/1
	3	0.5932	[0.5465, 0.6477]	40/0/0
	4	0.5373	[0.5073, 0.5708]	40/0/0
	6	0.4876	[0.4606, 0.5133]	40/0/0
	8	0.4302	[0.4091, 0.4545]	40/0/0
Act-SVD	2	0.7505	[0.6810, 0.8319]	37/0/3
	3	0.7422	[0.6815, 0.8059]	36/0/4
	4	0.7166	[0.6709, 0.7591]	39/0/1
	6	0.6789	[0.6320, 0.7196]	40/0/0
	8	0.6167	[0.5950, 0.6372]	40/0/0
Output-scored Act-SVD	2	0.8988	[0.8757, 0.9226]	36/0/4
	3	0.8772	[0.8546, 0.9016]	37/0/3
	4	0.8622	[0.8539, 0.8731]	39/0/1
	6	0.8203	[0.8019, 0.8338]	40/0/0
	8	0.7600	[0.7362, 0.7875]	40/0/0
AdaLoRA-score	2	0.8672	[0.8073, 0.9298]	33/0/7
	3	0.8448	[0.7922, 0.8994]	33/0/7
	4	0.8003	[0.7703, 0.8276]	38/0/2
	6	0.7328	[0.6988, 0.7634]	40/0/0
	8	0.6884	[0.6613, 0.7164]	40/0/0
IGU-LoRA-score	2	0.8548	[0.7966, 0.9118]	33/0/7
	3	0.8019	[0.7570, 0.8491]	38/0/2
	4	0.7358	[0.7037, 0.7704]	39/0/1
	6	0.7131	[0.6817, 0.7443]	40/0/0
	8	0.6621	[0.6395, 0.6864]	40/0/0

Table 10: Paired WAS/comparator KL ratio. Descriptive 95% intervals and W/T/L across 40 paired teacher cells per budget. Below one favors WAS.

Comparator	Rank	Estimate	95% interval	W/T/L
Uniform-SVD	2	0.3163	[0.2587, 0.3899]	40/0/0
	3	0.3355	[0.2952, 0.3901]	40/0/0
	4	0.2741	[0.2436, 0.3110]	40/0/0
	6	0.2358	[0.2118, 0.2597]	40/0/0
	8	0.1687	[0.1492, 0.1874]	40/0/0
PARA	2	0.3435	[0.2827, 0.4274]	40/0/0
	3	0.2893	[0.2424, 0.3560]	38/0/2
	4	0.2843	[0.2535, 0.3324]	40/0/0
	6	0.2419	[0.2061, 0.2861]	40/0/0
	8	0.2097	[0.1818, 0.2543]	40/0/0
Post-Squeeze	2	0.3163	[0.2587, 0.3899]	40/0/0
	3	0.3355	[0.2952, 0.3901]	40/0/0
	4	0.2741	[0.2436, 0.3110]	40/0/0
	6	0.2358	[0.2118, 0.2597]	40/0/0
	8	0.1687	[0.1492, 0.1874]	40/0/0
Act-SVD	2	0.5078	[0.4380, 0.5852]	40/0/0
	3	0.5121	[0.4482, 0.5911]	40/0/0
	4	0.4781	[0.4276, 0.5322]	39/0/1
	6	0.4408	[0.3920, 0.4911]	39/0/1
	8	0.3563	[0.3253, 0.3985]	39/0/1
Output-scored Act-SVD	2	0.8311	[0.7971, 0.8613]	34/0/6
	3	0.7781	[0.7454, 0.8186]	37/0/3
	4	0.7278	[0.6934, 0.7632]	38/0/2
	6	0.6827	[0.6497, 0.7165]	40/0/0
	8	0.5873	[0.5526, 0.6257]	40/0/0
AdaLoRA-score	2	0.7464	[0.6930, 0.7989]	36/0/4
	3	0.7282	[0.6804, 0.7742]	35/0/5
	4	0.6736	[0.6244, 0.7273]	37/0/3
	6	0.6014	[0.5678, 0.6367]	40/0/0
	8	0.5041	[0.4603, 0.5579]	39/0/1
IGU-LoRA-score	2	0.7491	[0.6882, 0.8155]	32/0/8
	3	0.7200	[0.6557, 0.7816]	37/0/3
	4	0.6238	[0.5738, 0.6779]	39/0/1
	6	0.5967	[0.5589, 0.6328]	40/0/0
	8	0.5032	[0.4721, 0.5397]	40/0/0

Table 11: Paired WAS/comparator accuracy difference (percentage points). Descriptive 95% intervals and W/T/L across 40 paired teacher cells per budget. Positive favors WAS.

Comparator	Rank	Estimate	95% interval	W/T/L
Uniform-SVD	2	0.7380	[0.2843, 1.2501]	27/3/10
	3	0.3663	[0.0755, 0.6985]	28/2/10
	4	0.2520	[0.0910, 0.4216]	29/3/8
	6	0.0953	[0.0234, 0.1617]	23/5/12
	8	0.0191	[-0.0968, 0.1299]	21/8/11
PARA	2	0.7049	[0.2979, 1.1654]	31/1/8
	3	0.4238	[0.1659, 0.6597]	31/2/7
	4	0.2915	[0.1130, 0.5079]	26/2/12
	6	0.0867	[-0.0892, 0.2190]	25/3/12
	8	0.0063	[-0.1195, 0.1298]	17/7/16
Post-Squeeze	2	0.7373	[0.2840, 1.2491]	27/3/10
	3	0.3661	[0.0749, 0.6980]	28/2/10
	4	0.2520	[0.0909, 0.4223]	29/3/8
	6	0.0953	[0.0234, 0.1617]	23/5/12
	8	0.0191	[-0.0968, 0.1299]	21/8/11
Act-SVD	2	0.4600	[0.1824, 0.7625]	27/3/10
	3	0.2213	[0.0840, 0.3580]	28/6/6
	4	0.1294	[-0.0086, 0.2656]	24/9/7
	6	0.0566	[0.0039, 0.1093]	18/6/16
	8	-0.0312	[-0.1195, 0.0546]	12/15/13
Output-scored Act-SVD	2	0.0937	[-0.0167, 0.1951]	21/8/11
	3	0.0946	[-0.1033, 0.2961]	23/3/14
	4	0.1137	[-0.0242, 0.2335]	24/4/12
	6	-0.0085	[-0.0822, 0.0480]	15/10/15
	8	-0.0414	[-0.0766, -0.0054]	12/13/15
AdaLoRA-score	2	0.3597	[0.2619, 0.5242]	23/4/13
	3	0.1998	[0.0330, 0.3537]	25/5/10
	4	0.0688	[-0.0071, 0.1530]	19/7/14
	6	-0.0047	[-0.0810, 0.0758]	14/6/20
	8	-0.0635	[-0.1270, -0.0032]	9/13/18
IGU-LoRA-score	2	0.2991	[0.1603, 0.4732]	27/2/11
	3	0.1965	[0.0525, 0.3400]	26/2/12
	4	0.0658	[-0.0102, 0.1449]	19/7/14
	6	0.0172	[-0.0556, 0.0885]	16/8/16
	8	-0.0378	[-0.1192, 0.0362]	13/13/14

E Protocol, Implementation, Reproducibility, and Cost

E.1 Data partitions and metrics

The five-task benchmark contains SST-2, QNLI, MRPC, RTE, and MNLI, with eight independently trained and development-selected rank-16 teachers per task. LoRA covers 24 query/value modules of dimension 768×768 in RoBERTa-base. All compression methods retain the same teacher and classifier, scaling $\gamma = 2$, and zero recovery steps. Total ranks are set to $K \in \{48, 72, 96, 144, 192\}$. The LoRA-factor parameter counts are $1536K$ versus 589,824 for the teacher. Equal total ranks therefore also match factor parameter counts.

Teacher-training, development, and calibration partitions are disjoint. Calibration/development sizes are 2,048/2,048 for SST-2, QNLI, and MNLI; 512/512 for MRPC; and 384/384 for RTE. Official validation is used for evaluation; both MNLI validation splits are retained. WAS and Act-SVD omit labels from statistics calls; supervised scoring controls use calibration labels. The evaluation includes 1,600 method–teacher–budget cells and 1,920 split records for the eight displayed methods.

For logits $Z, \widehat{Z} \in \mathbb{R}^{n \times C}$ and $H_C = I_C - \mathbf{1}\mathbf{1}^\top / C$, the primary fidelity metric is

$$e_{\mathrm{ctr}} = \frac{\|\widehat{Z} H_C - Z H_C\|_F}{\|Z H_C\|_F + 10^{-12}}. \tag{72}$$

We compute the KL divergence as $n^{-1} \sum_i \mathrm{KL}(\mathrm{softmax}(Z_i) \| \mathrm{softmax}(\widehat{Z}_i))$ at temperature one. Main accuracy first averages eight seeds per task and then gives each task weight $1/5$. MRPC uses raw label accuracy. MNLI uses the arithmetic mean of matched/mismatched accuracy within each teacher cell, not example-count pooling. Fidelity and official metrics use the same within-cell split convention.

E.2 Teacher training and selection

Teacher training and checkpoint selection follow the fixed configuration in Table 12, with development-only choices recorded for all 40 teachers.

Table 12: Teacher-training configuration and development-selection rules for the 40 fixed encoder teachers.

Field	Value	Selection/status
Backbone / tokenizer	RoBERTa-base / same model source	fixed
LoRA placement	query and value projections in 12 blocks; 24 modules	fixed
Source rank / alpha / scaling	16 / 32 / $\gamma = \alpha/r = 2$	fixed
Dropout / bias / head	LoRA dropout 0.05; bias none; classifier saved with adapter and retained unchanged during compression	fixed
Optimizer / decay	PyTorch AdamW / 0	fixed
Learning rates	$10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}$	selected per task/seed on development
Epochs	SST-2/QNLI/MNLI 3; MRPC 10; RTE 20	fixed
Batching	train 32; eval 64; accumulation 1	fixed
Length / schedule	maximum length 128; warmup fraction 0.06; linear schedule	fixed
Checkpoint rule	maximize development accuracy (MRPC: F1); ties: lower NLL, earlier step, lower learning rate	final split never used

E.3 Whitening and statistic collection

The implementation symmetrizes $\widehat{\Sigma}$ in float64 and uses

$$\lambda = \max\{c\ \mathrm{tr}(\widehat{\Sigma})/r, \epsilon_{64}\}, \qquad \widehat{\Sigma}_\lambda = \widehat{\Sigma} + \lambda I, \qquad c = 10^{-4}, \tag{73}$$

with eigenvalue clamping below ϵ_{64} before square and inverse-square roots. This numerical regularizer is distinct from the ridge-free theorem.

For encoders, the LoRA bottleneck activation tensor has shape $[B, T, r]$. Each raw class logit is summed across the calibration batch before differentiation; padded token rows are removed, and the valid gradient rows are concatenated into G, after which the implementation accumulates $G^\top G$. For N valid token rows and C classes, the gradient counter is NC, so the stored raw-coordinate statistic is $(NC)^{-1} \sum_{i,c} g_{i,c} g_{i,c}^\top$. Thus its whitened version is $\widehat{M}_H/C$ relative to the empirical form of (4), whereas the uncentered activation second moment divides by N. This common positive factor preserves each teacher's eigenspaces and global eigenvalue ordering; absolute residual values use the stored normalization. Raw logits define the primary GLUE scoring endpoint; centered relative logit error removes the common logit shift only when measuring fidelity. Token-row gradients remain context dependent through attention. The collector sums per-token gradient outer products; it does not form cross-token terms for a simultaneous sequence-level intervention. All compression-statistic collection is performed in evaluation mode, with dropout disabled.

For a hidden output of width d_H, the encoder uses masked Rademacher signs divided by $\sqrt{d_H}$ and averages the resulting Gram matrices over probes and valid token rows. It therefore estimates the endpoint Gram matrices divided by d_H. For integer $H \geq 1$, with source block index $b \in \{0, \ldots, L-1\}$ and L encoder blocks, the hidden-horizon collector selects the output of block $\min(b + H, L - 1)$; $H = 0$ denotes the immediate LoRA branch output, and $H = \mathrm{end}$ denotes the final-logit endpoint. The decoder uses vocabulary-centered, unscaled Rademacher signs on teacher-forced answer-token logits and divides accumulated Grams by the number of probes times valid input rows, with no additional vocabulary-width divisor (Section F). None of these estimators requires a materialized full Jacobian. Optional horizon mixtures use nonnegative weights; the encoder mixture utility first normalizes each raw-coordinate Gram collection by its total trace across modules. Such mixtures are separate from the primary pure raw-logit experiment.

E.4 Baseline implementation fidelity

Uniform-SVD truncates each learned update at uniform rank. PARA uses its compact QR–SVD core and native global spectral selection, without imposing a rank-one floor. Post-Squeeze uses the zero-recovery randomized construction with oversampling 10, two power iterations, and the teacher seed. Act-SVD is the exact local LoRA activation-aware control, not a full dense-model SVD-LLM reproduction. Act-SVD, Output-scored Act-SVD, WAS, and the supervised scoring controls use a rank-one floor. Output-scored Act-SVD is an information-matched constructed control. The adaptive scoring protocol is specified below.

Frozen-teacher adaptive scoring controls. AdaLoRA-score applies the sensitivity-times-uncertainty rule to exact SVD triplets of the frozen teacher. IGU-LoRA-score uses its Monte Carlo integrated-gradient scoring rule. Both select existing components while retaining their singular values, under the same total budgets and rank-one floor. No teacher weights or task heads are updated, and no recovery training is applied. WAS constructs a different unrestricted candidate subspace.

The implemented scoring protocol uses calibration labels, 25 scoring rounds of 16 minibatches, EMA coefficients 0.85, and an IGU discretization of 20 integration nodes (one sampled interior node plus the endpoint per minibatch; round-level score EMA). WAS instead collects downstream logit statistics without labels. These are supervised scoring adaptations, not information-matched unlabeled controls or complete reproductions of AdaLoRA/IGU adaptive training.

E.5 Complexity and measured cost

The completed cost benchmark is restricted to the same five tasks and 40 teachers. One-time scoring/statistics starts with an already loaded and warmed model and includes CPU/GPU work, transfers, SVD candidate preparation, and EMA updates where applicable. It excludes model loading, tokenization, disk writes, deletion audits, and final evaluation. Factor construction is measured separately: one warm-up followed by five timed repetitions per teacher/method/budget. Repetitions are first averaged within a cell, then seeds, tasks, and (for the main construction column) the five budgets receive equal weight. Statistics are not multiplied by the number of budgets. Output-scored Act-SVD shares WAS's statistics measurement rather than providing an independent repeat.

Hardware is NVIDIA A800-SXM4-80GB, one GPU and four CPU threads per job; Python 3.10.12, PyTorch 2.13.0+cu130, CUDA 13.0, Transformers 5.13.1, and PEFT 0.19.1. Model computation

is FP32; matmul TF32 is disabled and cuDNN TF32 enabled. Jobs may use different nodes, without exclusive-server or clock/power control. WAS/Act-SVD use batch 16. Supervised scoring uses $\min(64, \lfloor n_{\text{cal}}/16 \rfloor)$, namely 64 for SST-2/QNLI/MNLI, 32 for MRPC, and 24 for RTE, with dynamic padding and maximum length 128. Memory differences therefore also reflect batches and padding. Act-SVD uses the audited no-gradient forward collector; its timing and construction are from separate same-configuration jobs.

Table 13: One-time statistics/scoring and per-budget factor construction (seconds). Equal task/seed means; construction also averages the five target budgets. Statistics are collected once and reused.

Method	Statistics	Construction
Uniform-SVD	0.000	0.009
PARA	0.000	0.011
Post-Squeeze	0.000	0.036
Act-SVD	1.909	0.012
WAS	5.945	0.009
AdaLoRA-score	46.459	0.001
IGU-LoRA-score	89.391	0.001

Table 14: Statistics/scoring seconds per task, mean $\pm$ sample standard deviation over eight teachers.

Task	Act-SVD	WAS	AdaLoRA-score	IGU-LoRA-score
SST-2	1.85 ± 0.02	4.90 ± 0.05	33.13 ± 0.57	62.88 ± 0.59
QNLI	3.23 ± 0.01	9.69 ± 1.47	70.45 ± 0.52	136.41 ± 0.86
MRPC	0.75 ± 0.01	2.11 ± 0.04	30.74 ± 0.20	58.69 ± 0.28
RTE	0.78 ± 0.01	2.15 ± 0.02	33.85 ± 0.20	64.89 ± 0.25
MNLI	2.94 ± 0.02	10.88 ± 0.13	64.12 ± 0.34	124.08 ± 0.64

Table 15: Peak allocated GPU memory (MiB) per task, mean $\pm$ sample standard deviation over eight teachers.

Task	Act-SVD	WAS	AdaLoRA-score	IGU-LoRA-score
SST-2	523.95 ± 0.00	1190.43 ± 1.70	2207.19 ± 0.38	2207.19 ± 0.38
QNLI	559.77 ± 0.00	2025.31 ± 1.14	3937.82 ± 0.00	3937.82 ± 0.00
MRPC	547.33 ± 0.00	1589.32 ± 0.39	1896.56 ± 0.76	1896.48 ± 0.45
RTE	559.77 ± 0.00	2022.95 ± 0.00	1795.91 ± 0.00	1795.91 ± 0.00
MNLI	559.78 ± 0.00	1949.38 ± 0.99	3938.14 ± 0.35	3938.14 ± 0.35

The tables report PyTorch peak allocated GPU memory including the loaded model, not whole-process or reserved memory; uniform CPU peak memory was not measured. Per-task, per-seed, per-rank construction times and statistics memory are provided as CSV files. These timings characterize the implemented one-shot compression procedure, not optimized algorithmic bounds or inference latency.

E.6 Artifact regeneration and verification

The encoder records in `data/teacher_cells.csv` and `data/split_records.csv` retain teacher/prediction SHA-256 hashes, source identities, metrics, and rank patterns. `data/provenance.json` records source-file hashes; `data/verification.json` records coverage and the split-to-cell aggregation check. Stored prediction tensors for SST-2, QNLI, MRPC, and RTE do not contain explicit example identifiers; alignment was verified using official deterministic ordering, labels, and teacher predictions. This does not constitute stored-ID verification.

`scripts/build_encoder_tables.py` regenerates the seven consolidated encoder tables from `data/published_encoder_cells.json` and checks every displayed value against the CSV summaries at its published precision. It preserves the reported estimates, intervals, and W/T/L without recomputing the bootstrap. The vector accuracy figure and its mean/interval

data (`figures/five_task_accuracy.pdf`, `data/accuracy_curves.csv`) provide the plotted summary.

F Decoder-Only Experiments

F.1 Summary of decoder output preservation

The supplementary Qwen2.5-1.5B-Instruct study covers GSM8K (Cobbe et al., 2021) and ARC-Challenge (Clark et al., 2018), three fixed teachers per task, and average retained ranks $\bar{r} \in \{2, 4, 8\}$. In the matched-total-rank comparison, Table 16 shows lower centered error and KL in all 18 paired settings against every tested comparator. Relative to Act-SVD, error/KL fall by 23.7%/28.9%; relative to Output-scored Act-SVD, by 10.0%/12.8%. A complementary physical-parameter-matched comparison is reported in Table 20. These results support fidelity at a large-vocabulary endpoint.

Table 16: Decoder fidelity at matched total rank. Ratios are geometric means of WAS/comparator values across two tasks, three teachers per task, and three budgets. Parameter counts differ (Table 19).

Comparator	Error ratio	Wins	KL ratio	Wins
Uniform-SVD	0.739	18/18	0.718	18/18
PARA	0.589	18/18	0.575	18/18
H=0 Act-SVD	0.763	18/18	0.711	18/18
Output-scored H=0 Act-SVD	0.900	18/18	0.872	18/18

The comparisons match total rank. At $\bar{r} = 4$, mean adapter sizes over six teachers are 4.93M parameters for WAS, 4.75M for Output-scored Act-SVD, and 6.48M for Act-SVD (Table 19). Calibration uses 32 teacher-generated trajectories and eight probes; fidelity uses teacher forcing on reference answers.

F.2 Model, data, and teacher selection

The model is Qwen2.5-1.5B-Instruct with 28 blocks, hidden width 1,536, feed-forward width 8,960, and key/value projection width 256. Every block adapts query, key, value, attention output, gate, up, and down projections, giving 196 LoRA modules. Source rank is 16, $\alpha = 32$, dropout 0.05, and biases are frozen. The tokenizer uses the model's chat template, with one user message and a generation prompt. Runs use bfloat16.

Table 17: Decoder data and execution protocol. Calibration-pool size is distinguished from the 32 examples actually consumed for compression. Final data are the official test splits.

Item	GSM8K	ARC-Challenge
Teacher training examples	5,961	607
Held-out calibration pool	512	256
Compression examples used / teacher	32	32
Development examples	1,000	256
Final evaluation examples	1,319	1,172
Maximum input-plus-answer length	1,024	512
Greedy generation limit	256	8
Hutchinson probes	8	8

Teacher candidates use learning rates $\{10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$, one epoch capped at 1,000 optimizer steps, batch size one, and gradient accumulation eight. Selection maximizes development greedy exact match, breaking ties by answer-token accuracy and then lower NLL. GSM8K selects 10^{-4} for all three seeds; ARC-Challenge selects 5×10^{-4} for seeds 0 and 1 and 2×10^{-4} for seed 2. Final examples are not used in teacher selection.

Unlabeled calibration versus reference-answer evaluation. Each selected teacher greedily generates a response to held-out calibration prompts; these generated trajectories, rather than dataset

reference answers, supply the compression endpoint tokens. The collector consumes the first 32 trajectories in the deterministic file order, batch size one. They contribute 3,549, 3,750, and 3,689 GSM8K answer tokens for seeds 0–2, and 64 ARC-Challenge answer tokens per teacher, including EOS. GSM8K prompts request concise reasoning followed by a numeric answer; ARC prompts request only an option label.

Final fidelity is measured on the official test records' reference targets: the reference reasoning and answer for GSM8K and the correct option label for ARC. Both teacher and compressed models receive the same reference prefix. Prompt and padding labels are masked with -100; shifted answer labels define the evaluated token positions, including EOS. Thus teacher-forced output preservation is evaluated along reference trajectories, not along independently generated compressed-model trajectories. Greedy task scores are evaluated separately from the prompt alone. The template, masks, final records, and evaluation inputs are identical across methods within each teacher–budget comparison. Answer-bearing sequences are not truncated.

F.3 Endpoint estimation and fidelity metrics

For answer-token logits $Z_t, \widehat{Z}_t \in \mathbb{R}^V$, centered relative logit error accumulates $\sum_t \|H_V(\widehat{Z}_t - Z_t)\|_2^2$ and $\sum_t \|H_V Z_t\|_2^2$ over batches, then takes the square root of their ratio. Decoder KL sums $\mathrm{KL}(p_t \| \widehat{p}_t)$ over the vocabulary and answer tokens, then divides by the answer-token count. Both use temperature one. The activation second moment uses all non-padding input-token rows.

Eight Rademacher probes (Hutchinson, 1990) act on the 151,936-dimensional vocabulary, independently at each answer position, and are centered across vocabulary before the vector–Jacobian product. The random seed is teacher seed $+9173$, with batch/probe offset $1{,}000{,}003 \times$ batch index $+$ probe index. The collector divides its Gram by the number of probes times valid input rows. Within each run, $H{=}0$ Act-SVD, Output-scored $H{=}0$ Act-SVD, and WAS-LoRA consume the same activation statistics; the latter two also consume the same downstream Gram. The rank-matched study deterministically recomputes these statistics in a separate job for each budget; the parameter-matched study reuses each teacher's saved rank-2 statistics across its three budgets.

Table 18: Probe-count sanity check: Qwen2.5-1.5B-Instruct, GSM8K seed 0, 164,982 teacher-forced answer tokens. This cost-budgeted WAS feasibility check is separate from the matched-total-rank comparison.

Probes	Centered relative logit error	KL/token	Rank agreement vs. 32 probes
1	0.048536	0.001817	91.3%
2	0.048391	0.001797	90.8%
4	0.048457	0.001796	95.9%
8	0.048509	0.001807	92.9%
16	0.048422	0.001818	98.0%
32	0.048421	0.001810	100.0%

Probe-count sanity check. This one-teacher sanity check shows similar errors across the tested probe counts. The complete decoder comparison fixes eight probes for every task, teacher, and budget. No analogous probe-count ablation is needed for the primary encoder logit experiments: their low-dimensional class-logit Grams are accumulated exactly. We use stochastic probing for computationally practical estimation at the decoder's large-vocabulary endpoint.

F.4 Parameter-count audit under matched total rank

The rank-matched decoder comparisons use exact total ranks $K \in \{392, 784, 1{,}568\}$, corresponding to average retained ranks $\bar{r} \in \{2, 4, 8\}$. The three activation-/output-aware methods use minimum rank one; PARA uses its native zero floor, and Uniform-SVD uses the stated uniform rank. A query or attention-output direction costs 3,072 parameters, a key/value direction 1,792, and an MLP direction 10,496. Equal total rank therefore does not match physical parameter count. Table 19 gives all method means.

Table 19: Parameter-count audit under matched total rank. Mean physical LoRA parameters (millions) over six teachers, each retaining $196\bar{r}$ directions. Projection dimensions differ. The rank-16 teacher has 18.465M parameters.

Method	$\bar{r} = 2$	$\bar{r} = 4$	$\bar{r} = 8$
Uniform-SVD	2.308	4.616	9.232
PARA	3.805	7.456	12.485
$H{=}0$ Act-SVD	3.049	6.476	11.782
Output-scored $H{=}0$ Act-SVD	2.283	4.747	9.787
WAS-LoRA	2.336	4.932	10.194

Greedy task-score changes in the rank-matched study are mixed across tasks and budgets; the decoder study therefore focuses on teacher-function fidelity. Both comparisons evaluate compressed functions in source-rank modules with zero-padded factors.

F.5 PHYSICAL-PARAMETER-MATCHED DECODER FIDELITY

The complementary comparison fixes the physical LoRA-parameter budget to that of uniform ranks 2, 4, and 8: 2,308,096, 4,616,192, and 9,232,384 parameters, respectively. It uses the same six frozen teachers, final reference-answer trajectories, and zero-recovery teacher-forced fidelity metrics as the rank-matched study. For each teacher, all three budgets reuse the rank-2 job's saved eight-probe calibration statistics. PARA keeps its native zero-rank floor; $H{=}0$ Act-SVD, Output-scored $H{=}0$ Act-SVD, and WAS-LoRA retain at least one direction per module. For each method, a prefix-constrained parameter-budget allocation maximizes its own retained score. Total retained ranks may differ because projection widths differ. This comparison does not reevaluate greedy task scores.

Table 20: Decoder fidelity at matched physical LoRA-parameter budgets. Ratios are geometric means of WAS/comparator values across GSM8K and ARC-Challenge, three fixed teachers per task, and budgets corresponding to uniform ranks 2, 4, and 8. Values below one favor WAS.

Comparator	Error ratio	Wins	KL ratio	Wins
Uniform-SVD	0.783	16/18	0.780	17/18
PARA	0.650	18/18	0.682	18/18
$H{=}0$ Act-SVD	0.697	18/18	0.614	18/18
Output-scored $H{=}0$ Act-SVD	0.914	18/18	0.903	17/18

WAS has lower centered error than Uniform-SVD in 16/18 paired cells and lower KL in 17/18; against Output-scored $H{=}0$ Act-SVD, the corresponding counts are 18/18 and 17/18. Thus the rank-matched study's universal cell-wise advantage does not extend to every parameter-matched comparator and metric. All methods fit the stated caps; 89 of 90 method cells use the full budget, and the remaining cell leaves 256 parameters unused. Matching physical parameters does not fix the per-module ranks or numerical regularization.