SAD-LoRA: Spectral Alignment for Low-Rank Knowledge Distillation

Omer Tariq [1] **Syed Muhammad Raza** [1] **Jeongbae Son** [1]

Abstract

Distilling a fine-tuned teacher into a LoRA-adapted student is a standard recipe for parameter-efficient compression, but output-level KD does not explicitly control which rank-r weight subspace the adapter occupies. We propose **SAD-LoRA** (**S**pectral **A**lignment **D**istillation), which selects this subspace from the data-weighted student-space reference update $\Delta W_T \Sigma_x^{1/2}$ and maintains it during training via a differentiable principal-angle loss on $\mathrm{span}(B)$. We show that the data-weighted distillation error decomposes exactly into subspace misalignment, within-subspace coefficient mismatch, and irreducible rank residual; standard KD can affect the first term only indirectly through output gradients. On controlled synthetic problems with a flat teacher spectrum, SAD-LoRA reduces the subspace-misalignment term from 51% to nearly zero and lifts final subspace alignment from 0.49 to 1.00. On RoBERTa-large to RoBERTa-base distillation across six GLUE tasks, SAD-LoRA improves rank efficiency: at $r{=}4$, it matches or beats the strongest included spectral baseline on five of six tasks, and at $r{=}8$ it gives the best result on SST-2 and CoLA. Ablations identify subspace alignment as the load-bearing component, while coefficient matching is auxiliary.

1. Introduction

Parameter-efficient fine-tuning has become standard for adapting large pretrained models. Among existing approaches (Han et al., 2024), low-rank adaptation (LoRA) is particularly attractive because it freezes pretrained weights and learns a task-specific update $\Delta W = BA$, with $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, and $r \ll \min(d_{\text{out}}, d_{\text{in}})$ (Hu et al., 2022). In parallel, knowledge distillation (KD) compresses a stronger teacher into a smaller or capacity-constrained student (Hinton et al., 2015; Gou et al., 2021). Combining LoRA with KD is therefore natural for efficient compression: the student receives teacher supervision while only a low-rank update is trained (Azimi et al., 2024; Yang et al., 2024a).

This combination, however, leaves a structural question unresolved. Standard KD objectives operate on logits or intermediate activations, while LoRA constrains the student to a rank-r *weight* subspace. The KD loss can reduce output discrepancy without explicitly determining which r directions the adapter should occupy. Because the rank-r student cannot represent every direction of the teacher update, a misaligned column space of B wastes part of the rank budget on directions that are off-target with respect to the teacher's task-specific transformation. We refer to this failure mode as *subspace misalignment*: a distillation error component that is invisible to output-level KD but irreducible without rotating the adapter's column space.

Existing spectral LoRA methods address adjacent but distinct problems: they initialize or allocate adapters using the spectrum of the pretrained weight matrix (Meng et al., 2024; Wang et al., 2025; Büyükakyüz, 2024), activation-aware pretrained-weight decompositions (Yang et al., 2024b), or sensitivity-based subset selection (Zhong et al., 2024). These methods provide useful starting points, but they do not explicitly preserve a teacher-update subspace during distillation; after initialization, $\mathrm{span}(B)$ can still drift under task and KD gradients. The central issue is therefore not only how to initialize LoRA, but how to maintain the relevant low-rank subspace throughout training (Fig. 1).

This paper studies LoRA distillation from a spectral approximation perspective. Let $W_0^{(\ell)}$ be a pretrained student weight matrix and let $W_T^{(\ell)}$ denote a task-adapted reference weight in the same student parameter space, defining the student-space update $\Delta W_T^{(\ell)} = W_T^{(\ell)} - W_0^{(\ell)}$. A rank-$r$ LoRA student approximates this update by $B^{(\ell)}A^{(\ell)}$. Approximation quality depends not only on the singular values of $\Delta W_T^{(\ell)}$, but also on whether $\mathrm{span}(B^{(\ell)})$ aligns with the dominant output-side directions of the update under the data distribution. Because distillation is evaluated on data, the relevant spectral object is the data-weighted update $\widetilde{W}_T^{(\ell)} = \Delta W_T^{(\ell)} (\Sigma_x^{(\ell)})^{1/2}$, where $\Sigma_x^{(\ell)}$ is the layer-input

[1]Neubility Inc, Seoul, South Korea. Correspondence to: Omer Tariq <omer.tariq@neubility.co.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

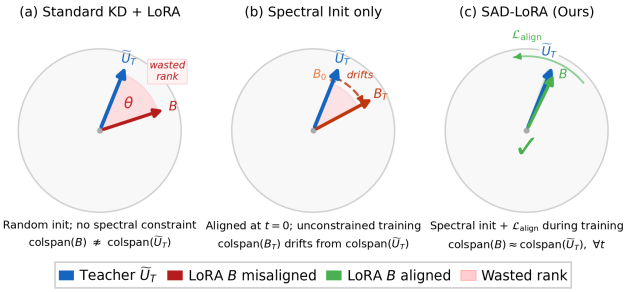

Figure 1. Three regimes for low-rank distillation. (a) Standard KD+LoRA may place $\mathrm{span}(B)$ away from the teacher subspace $\mathrm{span}(\widetilde{U}_{T,r})$, wasting rank budget. (b) Spectral initialization aligns the adapter at $t = 0$ but does not constrain later subspace drift. (c) SAD-LoRA combines spectral initialization with a continuous alignment loss $\mathcal{L}_{\text{align}}$ that encourages $\mathrm{span}(B) \approx \mathrm{span}(\widetilde{U}_{T,r})$ throughout training.

covariance.

We propose SAD-LoRA, which turns alignment with the leading left singular subspace of $\widetilde{W}_T^{(\ell)}$ into a training objective. SAD-LoRA augments KD with a Grassmannian principal-angle loss on $\mathrm{span}(B)$ and, optionally, a within-subspace coefficient-matching term. We show that the data-weighted distillation error decomposes exactly into subspace misalignment, coefficient mismatch, and rank residual; at fixed rank, only the first two are reducible by changing B and A. Standard KD can affect the subspace term only indirectly through output gradients, whereas SAD-LoRA targets it explicitly.

Contributions.

1. We formulate LoRA-based knowledge distillation as a data-weighted low-rank approximation problem and prove an exact three-term error decomposition (Theorem 4.1) that isolates subspace misalignment as a distinct, reducible source of error.

2. We identify the optimal rank-r adapter subspace as $\mathrm{span}(\widetilde{U}_{T,r}^{(\ell)})$, the leading left singular subspace of $\Delta W_T^{(\ell)}(\Sigma_x^{(\ell)})^{1/2}$, and derive a layerwise rank-sufficiency criterion from its spectral tail (Propositions 4.2–4.3).

3. We introduce SAD-LoRA, which uses a differentiable principal-angle alignment loss $\mathcal{L}_{\text{align}}$ and an optional coefficient-matching loss $\mathcal{L}_{\text{coeff}}$ to keep the LoRA adapter aligned with the data-weighted teacher-update subspace during distillation.

4. We validate SAD-LoRA on controlled synthetic problems, where the decomposition can be checked exactly, and on RoBERTa-large to RoBERTa-base distillation across six GLUE tasks, emphasizing rank efficiency and alignment-versus-coefficient ablations.

2. Related Work

PEFT and LoRA. Parameter-efficient fine-tuning adapts pretrained models by training small task-specific modules on top of frozen or near-frozen backbones (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Ben Zaken et al., 2022; Hu et al., 2022; Ding et al., 2023; Han et al., 2024). LoRA is widely used because it parameterizes a frozen-weight update as $\Delta W = BA$, preserves the merged architecture, and adds no inference latency (Hu et al., 2022). Subsequent variants improve LoRA through adaptive rank allocation, weight decomposition, quantization, parameter sharing, and modified parameterizations (Zhang et al., 2023; Valipour et al., 2023; Liu et al., 2024; Kopiczko et al., 2024; Renduchintala et al., 2024; Li et al., 2024a; Dettmers et al., 2023; Li et al., 2024b; Xu et al., 2024). SAD-LoRA keeps the same low-rank adapter form, but changes the training objective by explicitly aligning the adapter subspace with a data-weighted teacher-update target.

Spectral and activation-aware LoRA. A closer line of work exploits spectral structure to improve LoRA. AdaLoRA uses an SVD-like parameterization and prunes less important singular components (Zhang et al., 2023); PiSSA initializes adapters from principal singular components of the pretrained weight W_0 (Meng et al., 2024); MiLoRA uses minor components to preserve principal pretrained knowledge (Wang et al., 2025); and OLoRA uses QR-based orthonormal initialization for optimization stability (Büyükakyüz, 2024). These methods show that adapter subspace matters, but they decompose the pretrained weight or its compressed approximation, not the teacher's task-specific update. CorDA is closest to our data-weighted view: it orients pretrained-weight decomposition using input activation covariance from calibration data (Yang et al., 2024b). Activation-aware quantization and low-rank reconstruction use related covariance-aware principles (Yuan et al., 2023; Lin et al., 2024). SAD-LoRA differs in the decomposed object: CorDA uses $W_0\widehat{\Sigma}_x$, whereas SAD-LoRA uses the student-space teacher update $\Delta W_T\widehat{\Sigma}_x^{1/2}$. For KD, this distinction is central because the approximation target is the teacher's task update rather than the pretrained weight.

KD with LoRA students. KD transfers teacher behavior through labels, logits, or intermediate representations (Hinton et al., 2015; Gou et al., 2021). Recent methods combine KD with LoRA, including KD-LoRA for encoder benchmarks (Azimi et al., 2024), LLM-Neo for parameter-efficient LLM distillation (Yang et al., 2024a), and PC-LoRA for progressive compression (Hwang et al., 2024). These works establish KD+LoRA as an effective compression paradigm, but treat the adapter mainly as a parameter-efficient training channel; they do not specify which rank-r weight subspace should approximate the teacher update. Concurrent theory on low-rank KD provides

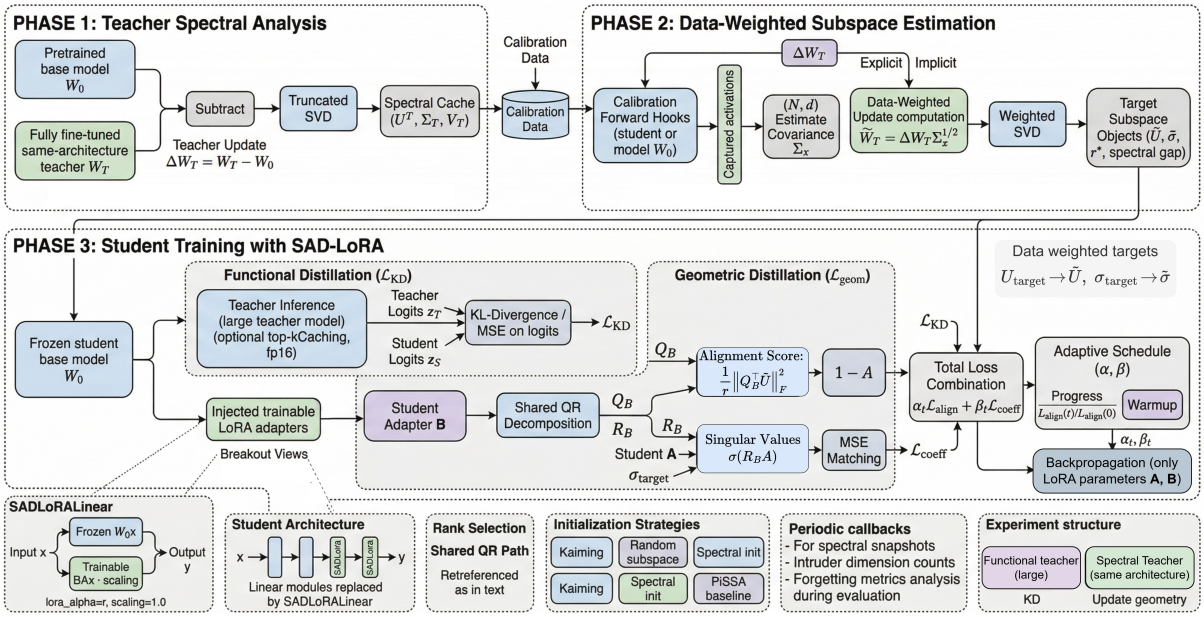

Figure 2. Overview of SAD-LoRA. Layerwise spectral targets $\widetilde{U}_{T,r}^{(\ell)}$ and $\widetilde{\sigma}_T^{(\ell)}$ are computed offline from a same-architecture reference teacher and calibration activations, then frozen. During training, the student receives output-level KD while the LoRA column space is geometrically aligned to the frozen targets. Only the LoRA adapters $\{B^{(\ell)}, A^{(\ell)}\}$ are optimized.

sample-complexity rank guarantees (Soarez et al., 2026); our criterion is complementary because it is layerwise, data-dependent, and based on the spectral tail of $\widetilde{W}_T$.

Subspace drift and intruder dimensions. Shuttleworth et al. (2024) show that LoRA trained with standard objectives can develop intruder dimensions: high-energy directions of BA that are orthogonal to both the pretrained spectrum and the teacher update and can drive forgetting. SAD-LoRA addresses the same geometric failure mode from the KD side: components of B outside the target data-weighted subspace increase $\mathcal{L}_{\text{align}}$, so any output-level gain must trade off against the alignment penalty.

3. Method

SAD-LoRA aligns the LoRA adapter with the data-weighted spectral subspace of a student-space teacher update. We first define the target update and its implicit SVD, then introduce the alignment and coefficient losses, the training objective, initialization, and computational cost. Fig. 2 gives an overview; full pseudocode is provided in Appendix A.

3.1. Setup and the Data-Weighted Teacher Update

Let $W_0^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote the pretrained weight matrix at layer ℓ. A LoRA student freezes $W_0^{(\ell)}$ and learns a rank-r_ℓ update

$$W_S^{(\ell)} = W_0^{(\ell)} + B^{(\ell)} A^{(\ell)}, \qquad r_\ell \ll \min(d_{\text{out}}, d_{\text{in}}), \quad (1)$$

with $B^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times r_\ell}$ and $A^{(\ell)} \in \mathbb{R}^{r_\ell \times d_{\text{in}}}$. The LoRA scaling factor is absorbed into $B^{(\ell)}A^{(\ell)}$. We suppress the layer index (ℓ) when context is clear. Let $W_T^{(\ell)}$ denote a task-adapted reference weight in the same parameter space as the student, and define the student-space reference update $\Delta W_T^{(\ell)} = W_T^{(\ell)} - W_0^{(\ell)}$. When the functional teacher providing distillation logits is larger than the student, this same-architecture reference model supplies only the weight-space spectral target; the larger teacher still supplies $\mathcal{L}_{\text{KD}}$. This separates functional supervision from the student-space geometry required for spectral alignment.

Output-level KD constrains predictions but not the rank-r_ℓ adapter subspace; SAD-LoRA derives this subspace from the teacher update under the downstream activation distribution. For layer input $x^{(\ell)}$ with covariance $\Sigma_x^{(\ell)} = \mathbb{E}[x^{(\ell)} x^{(\ell)\top}]$, the layerwise surrogate error satisfies

$$\begin{aligned} \mathcal{E}^{(\ell)} &= \mathbb{E}\Big[\big\|(\Delta W_T^{(\ell)} - B^{(\ell)}A^{(\ell)})\, x^{(\ell)}\big\|_2^2\Big] \\ &= \big\|(\Delta W_T^{(\ell)} - B^{(\ell)}A^{(\ell)})\,(\Sigma_x^{(\ell)})^{1/2}\big\|_F^2 \\ &= \big\|\widetilde{W}_T^{(\ell)} - B^{(\ell)}A^{(\ell)}(\Sigma_x^{(\ell)})^{1/2}\big\|_F^2, \end{aligned} \quad (2)$$

where the *data-weighted teacher update* is

$$\widetilde{W}_T^{(\ell)} = \Delta W_T^{(\ell)} (\Sigma_x^{(\ell)})^{1/2}. \quad (3)$$

The relevant spectral object is therefore $\widetilde{W}_T^{(\ell)}$, not the unweighted update alone. Let its compact SVD be

$$\widetilde{W}_T^{(\ell)} = \widetilde{U}_T^{(\ell)} \widetilde{\Sigma}_T^{(\ell)} \widetilde{V}_T^{(\ell)\top}, \qquad \widetilde{\sigma}_1^{(\ell)} \geq \widetilde{\sigma}_2^{(\ell)} \geq \cdots. \quad (4)$$

The leading left singular vectors $\widetilde{U}_{T,r}^{(\ell)} := \widetilde{U}_T^{(\ell)}[:, 1{:}r_\ell]$ define the target output-side subspace for $B^{(\ell)}$.

Implicit SVD. The covariance is never formed explicitly. Given centered calibration activations $X_c^{(\ell)} \in \mathbb{R}^{n \times d_{\text{in}}}$ collected from the student backbone on a small calibration set, define

$$Z^{(\ell)} = \tfrac{1}{\sqrt{n-1}}\, \Delta W_T^{(\ell)} X_c^{(\ell)\top} \in \mathbb{R}^{d_{\text{out}} \times n}. \tag{5}$$

Because $\widehat{\Sigma}_x^{(\ell)} = \frac{1}{n-1} X_c^{(\ell)\top} X_c^{(\ell)}$, we have

$$Z^{(\ell)} Z^{(\ell)\top} = \Delta W_T^{(\ell)} \widehat{\Sigma}_x^{(\ell)} \Delta W_T^{(\ell)\top} = \widetilde{W}_T^{(\ell)} \widetilde{W}_T^{(\ell)\top}, \tag{6}$$

so the left singular vectors of $Z^{(\ell)}$ are the empirical $\widetilde{U}_{T,r}^{(\ell)}$ and its singular values are the empirical $\widetilde{\sigma}_i^{(\ell)}$. This avoids forming the $d_{\text{in}} \times d_{\text{in}}$ covariance matrix when $n \ll d_{\text{in}}$. The targets $\widetilde{U}_{T,r}^{(\ell)}$ and $\widetilde{\sigma}_i^{(\ell)}$ are computed once and frozen for the rest of training.

3.2. Spectral Alignment and Coefficient Losses

The matrix $B^{(\ell)}$ alone determines the output-side subspace accessible to the LoRA adapter. Let $B^{(\ell)} = Q_B^{(\ell)} R_B^{(\ell)}$ be its thin QR decomposition. SAD-LoRA penalizes the principal-angle discrepancy between $\operatorname{span}(B^{(\ell)})$ and $\operatorname{span}(\widetilde{U}_{T,r}^{(\ell)})$:

$$\begin{aligned} \mathcal{L}_{\text{align}}^{(\ell)} &= 1 - \tfrac{1}{r_\ell} \big\| Q_B^{(\ell)\top} \widetilde{U}_{T,r}^{(\ell)} \big\|_F^2 \\ &= \tfrac{1}{r_\ell} \sum_{i=1}^{r_\ell} \sin^2 \theta_i^{(\ell)}, \end{aligned} \tag{7}$$

where $\theta_i^{(\ell)}$ are the principal angles between the two r_ℓ-dimensional subspaces. The loss is zero exactly when the adapter and target subspaces coincide and one when they are orthogonal.

Because $Q_B^{(\ell)}$ is orthonormal, the nonzero singular values of $B^{(\ell)} A^{(\ell)}$ equal those of $R_B^{(\ell)} A^{(\ell)}$, which is an $r_\ell \times d_{\text{in}}$ matrix and therefore much cheaper to decompose. We define a spectral coefficient-matching loss

$$\mathcal{L}_{\text{coeff}}^{(\ell)} = \frac{\big\| \sigma\Big(R_B^{(\ell)} A^{(\ell)}\Big) - \widetilde{\sigma}_{T,r}^{(\ell)} \big\|_2^2}{\big\| \widetilde{\sigma}_{T,r}^{(\ell)} \big\|_2^2 + \varepsilon}, \tag{8}$$

where $\widetilde{\sigma}_{T,r}^{(\ell)} = (\widetilde{\sigma}_1^{(\ell)}, \ldots, \widetilde{\sigma}_{r_\ell}^{(\ell)})$ and ε is a small constant. This is an unweighted proxy for the data-weighted coefficient objective analyzed in Section 4; it avoids applying $\Sigma_x^{1/2}$ at every step and is used only as an auxiliary regularizer.

3.3. Training Objective

For classification, we use temperature-scaled logit distillation; for regression, we use mean-squared error between teacher and student predictions. Let z_T, z_S denote teacher and student logits and τ the distillation temperature:

$$\mathcal{L}_{\text{KD}} = \tau^2\, \text{KL}(\text{softmax}(z_T/\tau) \,\|\, \text{softmax}(z_S/\tau))\,. \tag{9}$$

With supervised task loss $\mathcal{L}_{\text{task}}$, the SAD-LoRA objective is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{task}} &+ \lambda_{\text{KD}}\, \mathcal{L}_{\text{KD}} \\ &+ \alpha\, \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \mathcal{L}_{\text{align}}^{(\ell)} + \beta\, \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \mathcal{L}_{\text{coeff}}^{(\ell)}, \end{aligned} \tag{10}$$

where $\mathcal{S}$ is the set of adapted layers. The full objective uses both spectral terms ($\alpha, \beta > 0$); the alignment-only variant sets $\beta = 0$, the coefficient-only variant sets $\alpha = 0$. We keep the spectral weights α and β fixed throughout training; the optimizer uses the learning-rate warmup and decay described in Appendix C.

3.4. Spectral Initialization

SAD-LoRA initializes the adapter directly in the target subspace:

$$B^{(\ell)} = \widetilde{U}_{T,r}^{(\ell)}, \qquad A^{(\ell)} = 0. \tag{11}$$

Two consequences follow. First, $B^{(\ell)} A^{(\ell)} = 0$ at t=0, so the student function exactly reproduces the pretrained backbone. Second, $\mathcal{L}_{\text{align}}^{(\ell)}$ is identically zero at initialization, so $\mathcal{L}_{\text{KD}}$ supplies the dominant gradient signal from the first step and $\mathcal{L}_{\text{align}}$ acts as a regularizer that prevents drift rather than fighting the data fit. If the requested rank exceeds the effective target rank determined from the data-weighted spectral tail (see Proposition 4.3), we fill the remaining columns of $B^{(\ell)}$ with a random orthonormal complement obtained by Gram-Schmidt against $\widetilde{U}_{T,r}^{(\ell)}$. Zero-padding would leave $B^{(\ell)}$ rank-deficient and lock $\mathcal{L}_{\text{align}}$ above a hard floor regardless of training; orthonormal-complement padding guarantees both $\mathcal{L}_{\text{align}}^{(0)} = 0$ and full column rank.

3.5. Training Procedure and Cost

Spectral targets are computed once offline by forming the sketch $Z^{(\ell)}$ in Eq. 5 from calibration activations and registering $\widetilde{U}_{T,r}^{(\ell)}$ and $\widetilde{\sigma}_{T,r}^{(\ell)}$ as frozen buffers. Training then initializes $B^{(\ell)} = \widetilde{U}_{T,r}^{(\ell)}$, sets $A^{(\ell)} = 0$, and updates only the LoRA parameters using Eq. 10. The additional per-step cost is one thin QR of $B^{(\ell)}$ for $\mathcal{L}_{\text{align}}$ and, when $\mathcal{L}_{\text{coeff}}$ is used, one SVD of $R_B^{(\ell)} A^{(\ell)} \in \mathbb{R}^{r \times d_{\text{in}}}$, giving $O(r^2(d_{\text{out}} + d_{\text{in}}))$ per adapted layer. The targets are fixed after preprocessing, the backbone remains frozen, and no inference-time module is added. Full pseudocode is given in Appendix A.

4. Analysis

We analyze a single adapted layer and suppress the layer index. The data-weighted distillation error is

$$\mathcal{E}(B,A) = \left\|\widetilde{W}_T - BA\Sigma_x^{1/2}\right\|_F^2, \tag{12}$$

with $\widetilde{W}_T = \Delta W_T \Sigma_x^{1/2} = \widetilde{U}_T \widetilde{\Sigma}_T \widetilde{V}_T^\top$ its compact SVD.

4.1. Projector-Based Error Decomposition

Our first result is an exact projector-based decomposition of the data-weighted distillation error.

Theorem 4.1 (Distillation error decomposition)**.** *Let $B \in \mathbb{R}^{d_{\text{out}} \times r}$ have full column rank and let $P_B = B(B^\top B)^{-1}B^\top$ be the orthogonal projector onto $\operatorname{span}(B)$. Let $\widetilde{U}_{T,r}$ contain the top r left singular vectors of $\widetilde{W}_T$ and let $\mathcal{R} := \sum_{i>r} \widetilde{\sigma}_i^2$ be the rank-r tail energy. Then*

$$\mathcal{E}(B,A) \;=\; \mathcal{S}(B) \;+\; \mathcal{C}(B,A) \;+\; \mathcal{R}, \tag{13}$$

where

$$\mathcal{S}(B) := \left\|(I-P_B)\widetilde{W}_T\right\|_F^2 \;-\; \mathcal{R} \;\geq\; 0, \tag{14}$$

$$\mathcal{C}(B,A) := \left\|P_B\widetilde{W}_T - BA\Sigma_x^{1/2}\right\|_F^2 \;\geq\; 0. \tag{15}$$

The terms are interpretable as: (I) $\mathcal{S}(B)$ — subspace misalignment*, depending only on* $\operatorname{span}(B)$ *and irreducible without rotating B;* (II) $\mathcal{C}(B,A)$ — within-subspace coefficient mismatch*, eliminable by choosing A optimally given B; and* (III) $\mathcal{R}$ — rank residual*, fixed by the rank budget. Furthermore, if $\widetilde{\sigma}_r > \widetilde{\sigma}_{r+1}$, then $\mathcal{S}(B) = 0$ if and only if* $\operatorname{span}(B) = \operatorname{span}(\widetilde{U}_{T,r})$*. Under boundary degeneracy, $\mathcal{S}(B) = 0$ for any optimal rank-r left singular subspace within the degenerate eigenspace; SAD-LoRA uses the empirical SVD target $\widetilde{U}_{T,r}$ as the alignment reference.*

Sketch. Since $\operatorname{span}(BA\Sigma_x^{1/2}) \subseteq \operatorname{span}(B)$, Pythagoras gives

$$\mathcal{E}(B,A) = \|(I-P_B)\widetilde{W}_T\|_F^2 + \|P_B\widetilde{W}_T - BA\Sigma_x^{1/2}\|_F^2.$$

By Eckart–Young–Mirsky, the first term is lower-bounded by $\mathcal{R}$, with equality at a rank-r optimal left singular subspace, up to non-uniqueness under singular-value degeneracy. This yields Eq. 13; the second term is nonnegative by definition. Full proof in Appendix B.3. □

Theorem 4.1 clarifies why output-level KD can waste rank budget. Standard KD updates B and A through output gradients and can reduce coefficient error inside a misaligned subspace while leaving $\mathcal{S}$ large. SAD-LoRA addresses this by adding $\mathcal{L}_{\text{align}}$ (Section 3.2), which explicitly targets the subspace term.

$\mathcal{L}_{\text{align}}$ as a Grassmannian surrogate for $\mathcal{S}$. The alignment loss

$$\mathcal{L}_{\text{align}} \;=\; \tfrac{1}{r}\sum_{i=1}^{r} \sin^2\theta_i \tag{16}$$

is the normalized squared chordal distance on $\operatorname{Gr}(r, d_{\text{out}})$ between $\operatorname{span}(B)$ and the selected target subspace $\operatorname{span}(\widetilde{U}_{T,r})$. When the top-$r$ subspace is unique, $\mathcal{L}_{\text{align}} = 0$ and $\mathcal{S}(B) = 0$ have the same zero set. Under degeneracy, $\mathcal{S}$ may vanish for multiple optimal subspaces, while $\mathcal{L}_{\text{align}}$ anchors the adapter to the empirical SVD target used by SAD-LoRA. Thus $\mathcal{L}_{\text{align}}$ is not equal to $\mathcal{S}$; it is an unweighted, differentiable surrogate that is cheaper and more stable than estimating spectrally weighted angles directly.

Relation to $\mathcal{L}_{\text{coeff}}$. The principled, theory-aligned coefficient loss is the data-weighted form

$$\mathcal{L}_{\text{coeff}}^{\text{dw}} = \frac{\left\|\sigma\left(R_B A\Sigma_x^{1/2}\right) - \widetilde{\sigma}_{T,r}\right\|_2^2}{\left\|\widetilde{\sigma}_{T,r}\right\|_2^2 + \varepsilon}, \tag{17}$$

which corresponds to matching the within-subspace data-weighted spectrum implied by $\mathcal{C}$. Computing it requires applying $\Sigma_x^{1/2}$ (or equivalently right-multiplying by the calibration sketch X_c) at every step, which doubles the per-step cost of the coefficient term. Because our ablations show $\mathcal{L}_{\text{coeff}}$ to be auxiliary rather than load-bearing, we adopt the cheaper unweighted proxy $\mathcal{L}_{\text{coeff}} = \|\sigma(R_B A) - \widetilde{\sigma}_{T,r}\|_2^2 / (\|\widetilde{\sigma}_{T,r}\|_2^2 + \varepsilon)$ in our main experiments. The proxy equals $\mathcal{L}_{\text{coeff}}^{\text{dw}}$ when $\Sigma_x = I$; in the anisotropic case it is only an approximation to the data-weighted coefficient objective. We report results with the unweighted form throughout; preliminary runs with $\mathcal{L}_{\text{coeff}}^{\text{dw}}$ show no consistent advantage on GLUE.

4.2. Optimal Subspace and Rank Sufficiency

Proposition 4.2 (Data-weighted teacher subspace)**.** *Among all rank-r matrices M, an optimal minimizer of $\|\widetilde{W}_T - M\|_F^2$ has column space $\operatorname{span}(\widetilde{U}_{T,r})$ and residual $\sum_{i>r}\widetilde{\sigma}_i^2$. If $\widetilde{\sigma}_r > \widetilde{\sigma}_{r+1}$, this optimal column space is unique.*

Proof. Direct from Eckart–Young–Mirsky; see Appendix B.2. □

The relevant low-rank target is therefore not the unweighted teacher update, but the teacher update composed with $\Sigma_x^{1/2}$. Combined with Theorem 4.1, this identifies $\operatorname{span}(\widetilde{U}_{T,r})$ as an optimal minimizer of $\mathcal{S}(B)$, unique when $\widetilde{\sigma}_r > \widetilde{\sigma}_{r+1}$, and justifies $\mathcal{L}_{\text{align}}$ as its Grassmannian surrogate.

Proposition 4.3 (Rank sufficiency)**.** *For any $\epsilon > 0$, the minimum rank required to drive $\mathcal{E}(B,A) < \epsilon$ over all B, A*

is

$$r^*_\epsilon(\Delta W_T, \Sigma_x) = \min\left\{k : \sum_{i>k} \widetilde{\sigma}_i^2 < \epsilon\right\}. \tag{18}$$

Proof. At rank k, Eckart–Young–Mirsky gives the minimum attainable error $\sum_{i>k} \widetilde{\sigma}_i^2$. The stated criterion selects the smallest k for which this residual is below ϵ. See Appendix B.4. □

This is a layerwise, data-dependent rank criterion: layers with sharp spectral decay need fewer ranks, while layers with shallow decay require more. In the main experiments, we use a uniform rank r across layers to match prior LoRA baselines; Eq. 18 serves as a diagnostic and motivates future rank-allocation variants.

Implication for intruder dimensions. Any component of B outside the selected target subspace is penalized by $\mathcal{L}_{\text{align}}$. Under $\alpha > 0$, the optimizer must trade off any output-level gain from such directions against the geometric penalty, providing an explicit mechanism against intruder dimensions (Shuttleworth et al., 2024).

5. Experiments

We evaluate SAD-LoRA in two settings: controlled synthetic experiments that directly test the decomposition in Theorem 4.1, and RoBERTa-based GLUE distillation across ranks $r \in \{1, 2, 4, 8, 16\}$. We then ablate alignment and coefficient matching to identify which spectral component drives the gains.

5.1. Experimental Setup

Tasks and metrics. We evaluate on six GLUE tasks (Wang et al., 2018): SST-2, MRPC, STS-B, CoLA, QNLI, and RTE. We report accuracy for SST-2, QNLI, and RTE; F1 for MRPC; Spearman correlation for STS-B; and Matthews correlation (MCC) for CoLA. All GLUE results are mean and standard deviation over three seeds.

Models and adapted modules. The student is RoBERTa-base (Liu et al., 2020) with LoRA adapters on the query and value projections in each Transformer block. Functional distillation logits come from a RoBERTa-large teacher. Because SAD-LoRA requires teacher updates in the same parameter space as the student, spectral targets are computed from a same-architecture RoBERTa-base reference fine-tuned on the same task; the larger teacher continues to supply $\mathcal{L}_{\text{KD}}$. We evaluate $r \in \{1, 2, 4, 8, 16\}$.

Compared methods. All baselines are reproduced in the same training pipeline with matched optimizers and schedules: *KD-LoRA* (KL distillation with LoRA (Azimi et al., 2024)), *Logit-MSE* (replaces KL with MSE on logits), *NN-Init* (Neural Nuggets (Zhong et al., 2024), SVD initialization from ΔW_T without alignment objective), and *PiSSA-Init* (principal-singular-vector initialization from W_0 (Meng et al., 2024)). We additionally evaluate two SAD-LoRA ablations: *SAD-LoRA-Align* ($\beta = 0$, alignment only) and *SAD-LoRA-Coeff* ($\alpha = 0$, coefficient only). Hyperparameters are matched across methods; SAD-LoRA uses $\alpha = 0.2$, $\beta = 0.05$, $\tau = 4$, and calibration size $n = 1024$, with optimization details in Appendix C.

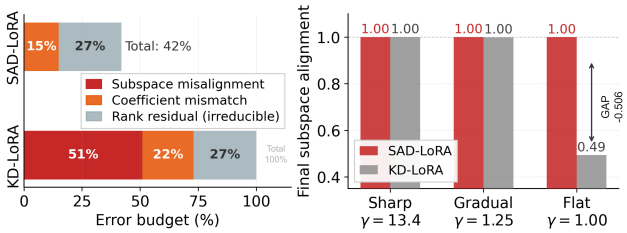

Figure 3. Controlled spectral validation. *Left:* under a flat teacher spectrum, KD-LoRA spends most of its error on subspace misalignment $\mathcal{S}$ (51%), whereas SAD-LoRA drives $\mathcal{S}$ nearly to zero, leaving the rank residual $\mathcal{R}$ (27%) and coefficient mismatch $\mathcal{C}$ (15%). *Right:* final alignment $\frac{1}{r}\|Q_B^\top \widetilde{U}_{T,r}\|_F^2$ across spectral gaps. Both methods align under sharp spectra, but only SAD-LoRA remains aligned in the flat case ($\gamma = 1$).

Table 1. Synthetic three-term error decomposition at $r = 4$ with flat teacher spectrum ($\gamma = 1$) and identity covariance.

Method	$\mathcal{S}$ subspace	$\mathcal{C}$ coeff.	$\mathcal{R}$ residual	Total
KD-LoRA	51%	22%	27%	100%
SAD-LoRA	$\approx$ **0**%	15%	27%	**42**%

5.2. Controlled Spectral Validation

The synthetic experiments test whether Theorem 4.1 predicts the actual error budget and whether $\mathcal{L}_{\text{align}}$ corrects the predicted failure mode. We generate teacher updates with controlled spectral gap $\gamma = \widetilde{\sigma}_r / \widetilde{\sigma}_{r+1}$, train rank-4 adapters, and compute $\mathcal{S}$, $\mathcal{C}$, and $\mathcal{R}$ exactly from the trained (B, A) and synthetic teacher.

Decomposition (Table 1). The flat-spectrum case is difficult for output-level KD because the teacher update provides weak directional preference. KD-LoRA spends most of its error budget on subspace misalignment ($\mathcal{S} = 51\%$), whereas SAD-LoRA drives $\mathcal{S}$ to nearly zero while leaving the irreducible rank residual $\mathcal{R}$ unchanged. The total error drops from 100% to 42%.

Alignment across spectral gaps. Fig. 3 (right) reports final subspace alignment $\frac{1}{r}\|Q_B^\top \widetilde{U}_{T,r}\|_F^2$ for sharp ($\gamma = 13.4$), gradual ($\gamma = 1.25$), and flat ($\gamma = 1$) teacher spectra. Both methods recover the dominant subspace under sharp spectra (1.00 each). As the spectral gap shrinks, KD-LoRA degrades sharply: its final alignment drops to 0.49 in the flat case, while SAD-LoRA remains at 1.00. This supports the

Table 2. GLUE results at ranks $r = 4$ and $r = 8$, reported as mean $\pm$ std over three seeds. SST-2, QNLI, and RTE use accuracy; MRPC uses F1; STS-B uses Spearman's ρ; and CoLA uses Matthews correlation. Best results/near-ties under seed variability are also bolded.

Method	SST-2	MRPC	STS-B	CoLA	QNLI	RTE
Rank $r = 4$						
KD-LoRA	93.3±0.4	87.8±0.9	0.789±0.033	0.426±0.013	91.3±0.2	65.0±1.6
Logit-MSE	93.8±0.2	87.2±0.2	0.789±0.033	0.423±0.013	90.8±0.6	59.0±3.7
NN-Init	93.4±0.4	**90.3±0.6**	**0.887±0.003**	0.555±0.008	**92.5±0.1**	69.2±1.0
PiSSA-Init	93.7±0.8	90.2±0.6	0.880±0.004	0.519±0.015	92.1±0.1	66.1±1.2
SAD-LoRA	94.3±0.2	89.9±0.3	0.884±0.002	**0.558±0.006**	92.1±0.0	69.9±1.3
SAD-LoRA-Align	94.1±0.1	90.1±0.2	**0.887±0.003**	0.553±0.003	**92.5±0.1**	**70.0±1.3**
SAD-LoRA-Coeff	**94.5±0.1**	90.0±0.4	0.885±0.002	0.539±0.009	92.2±0.0	63.1±6.8
Rank $r = 8$						
KD-LoRA	93.7±0.2	90.2±0.7	0.847±0.018	0.478±0.016	92.0±0.2	67.4±1.1
Logit-MSE	93.5±0.2	89.8±0.2	0.847±0.018	0.473±0.019	91.9±0.3	65.2±2.8
NN-Init	93.8±0.1	90.2±0.9	0.892±0.004	0.554±0.005	**92.8±0.2**	72.1±1.4
PiSSA-Init	94.0±0.5	**91.1±0.6**	0.887±0.006	0.538±0.015	92.3±0.1	68.8±2.1
SAD-LoRA	**94.5±0.2**	90.9±0.4	0.891±0.003	0.558±0.005	92.5±0.1	71.8±0.8
SAD-LoRA-Align	93.6±0.2	90.1±0.3	**0.893±0.003**	**0.562±0.006**	**92.8±0.1**	**72.2±0.6**
SAD-LoRA-Coeff	93.7±0.2	90.5±0.2	0.891±0.003	0.537±0.005	92.5±0.1	68.0±0.6

role of $\mathcal{L}_{\text{align}}$: it is most useful when the teacher spectrum provides weak directional preference.

5.3. Main GLUE Results

We report exact results at $r = 4$ and $r = 8$, and show rank trajectories for $r \in \{1, 2, 4, 8, 16\}$.

Rank-$4/8$ results (Table 2). At $r = 4$, SAD-LoRA-family variants are strongest or competitive on five of six tasks. SAD-LoRA-Coeff gives the best SST-2 score (94.5), full SAD-LoRA gives the best CoLA score (0.558), and SAD-LoRA-Align matches NN-Init on STS-B (0.887) and QNLI (92.5) while giving the best RTE score (70.0). MRPC is the exception: NN-Init remains strongest (90.3), with SAD-LoRA variants within 0.4 F1. Relative to KD-LoRA, full SAD-LoRA improves STS-B from 0.789 to 0.884, CoLA from 0.426 to 0.558, and RTE from 65.0 to 69.9. Notably, SAD-LoRA at $r = 4$ already exceeds KD-LoRA at $r = 8$ on STS-B and CoLA, indicating that subspace selection can compensate for additional rank. At $r = 8$, margins shrink on saturated tasks, but the same pattern remains. Full SAD-LoRA gives the best SST-2 score (94.5), while SAD-LoRA-Align is strongest on STS-B (0.893), CoLA (0.562), and RTE (72.2), and ties NN-Init on QNLI (92.8). PiSSA-Init remains best on MRPC (91.1), with full SAD-LoRA close behind (90.9). The largest gains over KD-LoRA again occur on STS-B and CoLA, supporting the hypothesis that data-weighted teacher-update directions matter most when the adapter must choose a small task-relevant subspace.

Rank trajectories (Fig. 4). STS-B gives the clearest rank-efficiency signal: at $r = 1$, Logit-MSE collapses to -0.016 Spearman while SAD-LoRA already reaches 0.85, and by $r = 4$ SAD-LoRA-Align and NN-Init match the best-rank KD baselines. CoLA shows a similar low-rank gain, whereas SST-2 and QNLI are nearly saturated and MRPC favors PiSSA-Init at moderate rank.

5.4. Ablations and Rank Efficiency

The ablations isolate three effects: output-level KD (KD-LoRA, Logit-MSE), spectral initialization without training-time alignment (NN-Init, PiSSA-Init), and SAD-LoRA's two spectral terms separately.

Subspace alignment is load-bearing. Removing $\mathcal{L}_{\text{align}}$ (SAD-LoRA-Coeff) is more damaging than removing $\mathcal{L}_{\text{coeff}}$. At $r = 8$, dropping $\mathcal{L}_{\text{align}}$ costs 0.025 MCC on CoLA ($0.562 \to 0.537$) and 4.2 accuracy points on RTE ($72.2 \to 68.0$). The variance of SAD-LoRA-Coeff is also larger on small-data tasks (± 6.8 on RTE at $r = 4$). This is consistent with Theorem 4.1: $\mathcal{L}_{\text{coeff}}$ targets $\mathcal{C}$, which is bounded by the within-subspace projection of $\widetilde{W}_T$ and is therefore meaningful only after $\mathcal{S}$ has been driven small.

Coefficient matching is auxiliary. Removing $\mathcal{L}_{\text{coeff}}$ improves or matches the full objective on STS-B, CoLA, QNLI, and RTE while the full objective is stronger on SST-2 and MRPC. This suggests that exact singular-value matching is more sensitive to weighting than directional alignment and is best used as a regularizer rather than the central mechanism.

Per-task summary at $r = 8$ (Fig. 5). SAD-LoRA variants improve over the best non-SAD baseline on SST-2 ($+0.5$), STS-B ($+0.1$), CoLA ($+0.8$), and RTE ($+0.1$); tie on QNLI; and remain 0.2 F1 below PiSSA-Init on MRPC.

Synthesis. The experiments support two conclusions. (i) The synthetic decomposition confirms Theorem 4.1: standard KD's distillation error is dominated by subspace misalignment under flat spectra, and SAD-LoRA removes this term. (ii) GLUE results show that this mechanism translates into downstream gains in low-rank, spectrally sensitive

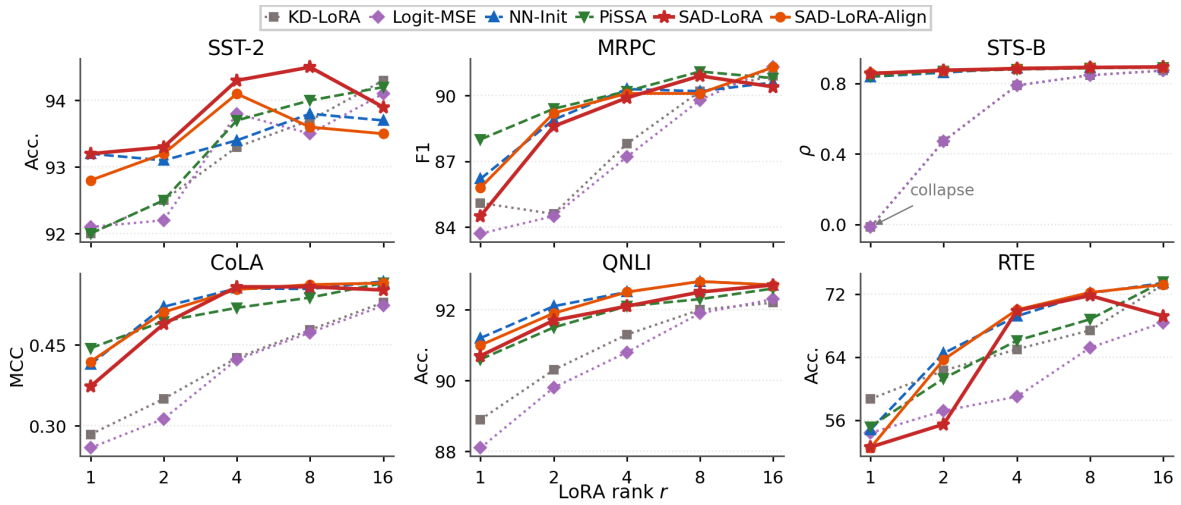

Figure 4. Rank efficiency on GLUE. Mean validation performance over three seeds as the LoRA rank varies from 1 to 16. SAD-LoRA variants are most effective in low-rank regimes. STS-B and CoLA show the clearest separation; SST-2 and QNLI are near-saturated; on MRPC, PiSSA-Init remains strongest at moderate rank. Logit-MSE collapses on STS-B at $r = 1$ (Spearman near zero), highlighting the value of subspace-aware initialization for rank-limited correlation tasks.

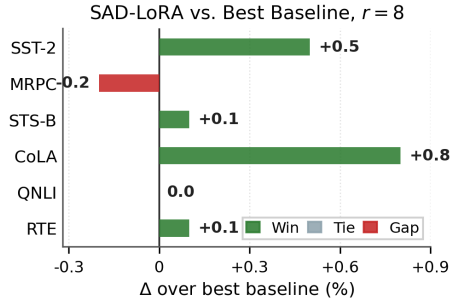

Figure 5. Best SAD-LoRA-family result minus best non-SAD baseline at $r = 8$. Positive values indicate an advantage for SAD-LoRA. Correlation and MCC differences are shown in percentage-point units.

regimes, especially on STS-B and CoLA.

6. Discussion and Limitations

SAD-LoRA's gains come primarily from controlling the adapter subspace during training, not merely from spectral initialization. This distinction is important because PiSSA-Init and NN-Init already use SVD-based initialization, yet they do not keep $\mathrm{span}(B)$ tied to the data-weighted teacher-update subspace. The synthetic decomposition shows that this geometric target removes the subspace-misalignment term that KD-LoRA does not explicitly penalize, while the GLUE results show corresponding rank-efficiency gains: at $r = 4$, SAD-LoRA already matches or exceeds KD-LoRA at $r = 8$ on STS-B and CoLA. The ablations further indicate that $\mathcal{L}_{\text{align}}$ is the central mechanism. Coefficient matching is useful as an auxiliary regularizer—especially on tasks such as SST-2 and MRPC—but it is less robust than directional alignment and should be treated as optional rather than load-bearing.

Limitations. SAD-LoRA requires a teacher update in the same weight space as the student adapter. When the functional teacher is larger than the student, we therefore use a same-architecture fine-tuned reference model to construct the spectral targets, adding one offline model to the pipeline. The method also requires a calibration pass to compute layer-wise targets, although this cost is incurred once and adds no inference overhead. Empirically, the gains are not uniform: MRPC remains favorable to PiSSA-Init at moderate rank, suggesting that pretrained-weight principal structure can be preferable for some semantic-matching tasks. Finally, our experiments focus on encoder-model distillation on GLUE; extending teacher-update alignment to decoder-only LLMs and generation tasks remains future work.

7. Conclusion

We introduced SAD-LoRA, a spectral alignment framework for low-rank knowledge distillation. SAD-LoRA derives a layerwise target subspace from the data-weighted student-space reference update $\Delta W_T \Sigma_x^{1/2}$ and regularizes the LoRA adapter to remain aligned with this subspace during training. The analysis decomposes the data-weighted distillation error into subspace misalignment, coefficient mismatch, and rank residual, showing why output-level KD alone can waste rank budget in a misaligned adapter subspace. Synthetic experiments verify this mechanism directly, and RoBERTa-based GLUE distillation shows measurable rank-efficiency gains, especially on STS-B and CoLA. Ablations indicate that principal-angle subspace alignment is the central component, while coefficient matching is useful but auxiliary.

Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00443391, Korea-Japan Joint International Research on Deep Learning-based Autonomous Mobility Control Technology for Autonomous Mobile Robots).

References

Azimi, R., Rishav, R., Teichmann, M., and Ebrahimi Kahou, S. KD-LoRA: A hybrid approach to efficient fine-tuning with LoRA and knowledge distillation. In Rezagholizadeh, M., Passban, P., Samiee, S., Partovi Nia, V., Cheng, Y., Deng, Y., Liu, Q., and Chen, B. (eds.), *Proceedings of The 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, volume 262 of *Proceedings of Machine Learning Research*, pp. 73–80. PMLR, 2024.

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.1.

Büyükakyüz, K. OLoRA: Orthonormal low-rank adaptation of large language models, 2024.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., and Sun, M. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235, 2023. doi: 10.1038/s42256-023-00626-4.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. doi: 10.1007/s11263-021-01453-z.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. NIPS Deep Learning and Representation Learning Workshop.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Hwang, I., Park, H., Lee, Y., Yang, J., and Maeng, S. PC-LoRA: Low-rank adaptation for progressive model compression with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2024.

Kopiczko, D. J., Blankevoort, T., and Asano, Y. M. VeRA: Vector-based random matrix adaptation. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353.

Li, Y., Han, S., and Ji, S. VB-LoRA: Extreme parameter efficient fine-tuning with vector banks. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Li, Y., Yu, Y., Liang, C., He, P., Karampatziakis, N., Chen, W., and Zhao, T. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024b.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

Meng, F., Wang, Z., and Zhang, M. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Renduchintala, A., Konuk, T., and Kuchaiev, O. Tied-LoRA: Enhancing parameter efficiency of LoRA with weight tying. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8694–8705. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.481.

Shuttleworth, R., Andreas, J., Torralba, A., and Sharma, P. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024.

Soarez, A. R., Kim, D., Costa, M., and Torre, A. Demystifying low-rank knowledge distillation in large language models: Convergence, generalization, and information-theoretic guarantees. *arXiv preprint arXiv:2603.22355*, 2026. doi: 10.48550/arXiv.2603.22355.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., and Ghodsi, A. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.239.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Wang, H., Li, Y., Wang, S., Chen, G., and Chen, Y. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4823–4836, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.248. URL https://aclanthology.org/2025.naacl-long.248/.

Xu, Y., Xie, L., Gu, X., Chen, X., Chang, H., Zhang, H., Chen, Z., Zhang, X., and Tian, Q. QA-LoRA: Quantization-aware low-rank adaptation of large language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Yang, R., Wu, T., Wang, J., Hu, P., Wu, Y.-C., Wong, N., and Yang, Y. LLM-Neo: Parameter efficient knowledge distillation for large language models, 2024a.

Yang, Y., Li, X., Zhou, Z., Song, S. L., Wu, J., Nie, L., and Ghanem, B. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Yuan, Z., Shang, Y., Song, Y., Yang, D., Wu, Q., Yan, Y., and Sun, G. ASVD: Activation-aware singular value decomposition for compressing large language models, 2023.

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Zhong, M., An, C., Chen, W., Han, J., and He, P. Seeking neural nuggets: Knowledge transfer in large language models from a parametric perspective. In *International Conference on Learning Representations*, volume 2024, pp. 25576–25596, 2024.

A. SAD-LoRA Pseudocode

Algorithm 1 SAD-LoRA Training

Input: Pretrained student weights $\{W_0^{(\ell)}\}$; same-architecture reference teacher $\{W_T^{\mathrm{ref},(\ell)}\}$ in the student parameter space; functional teacher f_T supplying logits; training data $\mathcal{D}$; calibration set $\mathcal{D}_{\mathrm{cal}}$; adapted layers $\mathcal{S}$; ranks $\{r_\ell\}_{\ell\in\mathcal{S}}$; hyperparameters $\alpha, \beta, \tau, \lambda_{\mathrm{KD}}$.

Offline spectral target construction.
for each adapted layer $\ell \in \mathcal{S}$ **do**
  Compute the student-space reference update $\Delta W_T^{(\ell)} \leftarrow W_T^{\mathrm{ref},(\ell)} - W_0^{(\ell)}$.
  Collect centered calibration activations $X_c^{(\ell)} \in \mathbb{R}^{n\times d_{\mathrm{in}}}$ from the frozen student backbone.
  Form the sketch $Z^{(\ell)} \leftarrow \frac{1}{\sqrt{n-1}} \Delta W_T^{(\ell)} X_c^{(\ell)\top}$.
  Compute truncated SVD of $Z^{(\ell)}$ to obtain $\widetilde{U}_{T,r}^{(\ell)}$ and $\widetilde{\sigma}_{T,r}^{(\ell)}$.
  Register $\widetilde{U}_{T,r}^{(\ell)}$ and $\widetilde{\sigma}_{T,r}^{(\ell)}$ as frozen buffers.
end for

Spectral initialization.
for each adapted layer $\ell \in \mathcal{S}$ **do**
  Initialize $B^{(\ell)} \leftarrow \widetilde{U}_{T,r}^{(\ell)}$ and $A^{(\ell)} \leftarrow 0$.
  if r_ℓ exceeds the effective spectral rank **then**
    Pad the remaining columns of $B^{(\ell)}$ with a random orthonormal complement.
  end if
end for

Training.
for each minibatch $(x, y) \sim \mathcal{D}$ **do**
  $z_T \leftarrow f_T(x)$ *no gradient*
  $z_S \leftarrow f_S(x)$
  $\mathcal{L}_{\mathrm{task}} \leftarrow \mathrm{TaskLoss}(z_S, y)$
  $\mathcal{L}_{\mathrm{KD}} \leftarrow \tau^2\, \mathrm{KL}(\mathrm{softmax}(z_T/\tau) \,\|\, \mathrm{softmax}(z_S/\tau))$
  for each adapted layer $\ell \in \mathcal{S}$ **do**
    $Q_B^{(\ell)}, R_B^{(\ell)} \leftarrow \mathrm{QR}(B^{(\ell)})$
    $\mathcal{L}_{\mathrm{align}}^{(\ell)} \leftarrow 1 - \frac{1}{r_\ell} \left\| Q_B^{(\ell)\top} \widetilde{U}_{T,r}^{(\ell)} \right\|_F^2$
    $\hat{\sigma}^{(\ell)} \leftarrow \mathrm{svdvals}\left(R_B^{(\ell)} A^{(\ell)} \right)$
    $\mathcal{L}_{\mathrm{coeff}}^{(\ell)} \leftarrow \frac{\left\| \hat{\sigma}^{(\ell)} - \widetilde{\sigma}_{T,r}^{(\ell)} \right\|_2^2}{\left\| \widetilde{\sigma}_{T,r}^{(\ell)} \right\|_2^2 + \varepsilon}$
  end for
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{KD}} \mathcal{L}_{\mathrm{KD}} + \frac{\alpha}{|\mathcal{S}|} \sum_{\ell\in\mathcal{S}} \mathcal{L}_{\mathrm{align}}^{(\ell)} + \frac{\beta}{|\mathcal{S}|} \sum_{\ell\in\mathcal{S}} \mathcal{L}_{\mathrm{coeff}}^{(\ell)}$
  Backpropagate $\mathcal{L}$ and update only $\{B^{(\ell)}, A^{(\ell)}\}_{\ell\in\mathcal{S}}$ with AdamW.
end for

B. Proofs of Theoretical Results

This appendix collects the full proofs of the analytical results in Section 4. We use the same notation: $\widetilde{W}_T = \Delta W_T \Sigma_x^{1/2} \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$, with compact SVD $\widetilde{W}_T = \widetilde{U}_T \widetilde{\Sigma}_T \widetilde{V}_T^\top$ and singular values $\widetilde{\sigma}_1 \geq \widetilde{\sigma}_2 \geq \cdots \geq 0$. We write $\widetilde{U}_{T,r} = \widetilde{U}_T[:, 1{:}r]$ for the matrix of top-r left singular vectors and $\widetilde{W}_T^{(r)} = \widetilde{U}_{T,r} \widetilde{\Sigma}_{T,r} \widetilde{V}_{T,r}^\top$ for the rank-r truncation.

B.1. Useful Identities

We collect three elementary facts used throughout.

(F1) Pythagoras for the Frobenius norm. For any matrix $X \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$ and any orthogonal projector $P \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{out}}}$,

$$\|X\|_F^2 = \|PX\|_F^2 + \|(I-P)X\|_F^2,$$

because P and $I - P$ have orthogonal column ranges and Frobenius inner products of column-orthogonal matrices vanish.

(F2) Eckart–Young–Mirsky. The minimum of $\|\widetilde{W}_T - M\|_F^2$ over all rank-r matrices M is $\sum_{i>r} \widetilde{\sigma}_i^2$, attained by $\widetilde{W}_T^{(r)} = \widetilde{U}_{T,r} \widetilde{\Sigma}_{T,r} \widetilde{V}_{T,r}^\top$, whose column space is $\mathrm{span}(\widetilde{U}_{T,r})$.

(F3) Projection of a matrix onto a subspace via P_B. For B with full column rank and $P_B = B(B^\top B)^{-1}B^\top$,

$$\|P_B X\|_F^2 = \mathrm{tr}(X^\top P_B X) = \sum_i \|P_B x_i\|_2^2$$

where x_i denotes the i-th column of X.

B.2. Proof of Proposition 4.2

Proof. Direct application of Eckart–Young–Mirsky (F2) to $\widetilde{W}_T$. The minimum over rank-r matrices of $\|\widetilde{W}_T - M\|_F^2$ equals $\sum_{i>r} \widetilde{\sigma}_i^2$ and is attained by $\widetilde{W}_T^{(r)} = \widetilde{U}_{T,r}\widetilde{\Sigma}_{T,r}\widetilde{V}_{T,r}^\top$. Its column space is $\mathrm{span}(\widetilde{U}_{T,r})$. □

B.3. Proof of Theorem 4.1

Proof. Let $M = BA\Sigma_x^{1/2} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and let $P_B = B(B^\top B)^{-1}B^\top$ be the orthogonal projector onto $\mathrm{span}(B)$. Since $\mathrm{span}(M) \subseteq \mathrm{span}(B)$, we have $P_B M = M$, hence $(I - P_B)M = 0$.

Step 1: Pythagorean splitting. Apply (F1) to $X = \widetilde{W}_T - M$:

$$\begin{aligned}\mathcal{E}(B, A) \;=\; \|\widetilde{W}_T - M\|_F^2 &= \|P_B(\widetilde{W}_T - M)\|_F^2 + \|(I - P_B)(\widetilde{W}_T - M)\|_F^2 \\ &= \|P_B\widetilde{W}_T - M\|_F^2 + \|(I - P_B)\widetilde{W}_T\|_F^2,\end{aligned}$$

where the second equality uses $P_B M = M$ and $(I - P_B)M = 0$. This is the exact two-term projector decomposition.

Step 2: Identifying the rank residual. Define $\mathcal{R} = \sum_{i>r} \widetilde{\sigma}_i^2$. By (F2), for any rank-$r$ orthogonal projector P,

$$\|(I - P)\widetilde{W}_T\|_F^2 \;=\; \|\widetilde{W}_T\|_F^2 - \|P\widetilde{W}_T\|_F^2 \;=\; \sum_{i\geq 1} \widetilde{\sigma}_i^2 - \|P\widetilde{W}_T\|_F^2.$$

The maximum of $\|P\widetilde{W}_T\|_F^2$ over rank-r projectors is $\sum_{i=1}^r \widetilde{\sigma}_i^2$, achieved iff P projects onto $\mathrm{span}(\widetilde{U}_{T,r})$. Therefore

$$\|(I - P_B)\widetilde{W}_T\|_F^2 \;\geq\; \mathcal{R},$$

with equality if and only if $\mathrm{span}(B) = \mathrm{span}(\widetilde{U}_{T,r})$.

Step 3: Defining $\mathcal{S}(B)$. Set

$$\mathcal{S}(B) \;:=\; \|(I - P_B)\widetilde{W}_T\|_F^2 - \mathcal{R} \;\geq\; 0.$$

By Step 2, $\mathcal{S}(B) = 0$ iff $\mathrm{span}(B) = \mathrm{span}(\widetilde{U}_{T,r})$.

Step 4: Defining $\mathcal{C}(B, A)$. Set

$$\mathcal{C}(B, A) \;:=\; \|P_B\widetilde{W}_T - BA\Sigma_x^{1/2}\|_F^2 \;\geq\; 0,$$

which is nonnegative by definition.

Step 5: Combining. Substituting back into Step 1:

$$\mathcal{E}(B, A) \;=\; \|P_B\widetilde{W}_T - BA\Sigma_x^{1/2}\|_F^2 + \|(I - P_B)\widetilde{W}_T\|_F^2 \;=\; \mathcal{C}(B, A) + \mathcal{S}(B) + \mathcal{R}.$$

This is Eq. 13. The term $\mathcal{R}$ depends only on the rank budget; $\mathcal{S}(B)$ depends only on $\mathrm{span}(B)$; $\mathcal{C}(B, A)$ depends on both B and A.

Optimality of A given B. For any fixed B, $\mathcal{C}(B, A)$ is a quadratic in A minimized at the (pseudo-)inverse solution. Specifically, when Σ_x is full rank,

$$A^*(B) \;=\; (R_B^\top R_B)^{-1} R_B^\top Q_B^\top \widetilde{W}_T\, (\Sigma_x^{1/2})^{-1},$$

so $\mathcal{C}(B, A^*) = 0$. When $\widetilde{\sigma}_r > \widetilde{\sigma}_{r+1}$, $\mathcal{S}(B) = 0$ characterizes the unique optimal output subspace at fixed rank r. Under degeneracy, the optimal subspace is not unique, and $\widetilde{U}_{T,r}$ denotes the empirical SVD target selected for alignment. □

B.4. Proof of Proposition 4.3

Proof. Fix $\epsilon > 0$. By Theorem 4.1, $\mathcal{E}(B, A) \geq \mathcal{R} = \sum_{i>r} \widetilde{\sigma}_i^2$ at any rank r, with equality attained by $B^* = \widetilde{U}_{T,r}$ and the corresponding optimal A^* from Step 5 of Appendix B.3. Therefore

$$\min_{B,A:\,\mathrm{rank}(B)\leq k} \mathcal{E}(B, A) \;=\; \sum_{i>k} \widetilde{\sigma}_i^2.$$

The smallest k for which this minimum drops below ϵ is exactly

$$r_\epsilon^*(\Delta W_T, \Sigma_x) \;=\; \min\left\{ k \in \mathbb{N} : \sum_{i>k} \widetilde{\sigma}_i^2 < \epsilon \right\},$$

which is Eq. 18. □

B.5. Surrogate Relationship: $\mathcal{L}_{\text{align}}$ and $\mathcal{S}(B)$

We record the precise sense in which $\mathcal{L}_{\text{align}}$ is a surrogate for $\mathcal{S}(B)$.

Lemma B.1 (Zero set)**.** *$\mathcal{L}_{\text{align}}(B) = 0$ if and only if $\mathrm{span}(B) = \mathrm{span}(\widetilde{U}_{T,r})$, if and only if $\mathcal{S}(B) = 0$.*

Proof. Let $B = Q_B R_B$ be the thin QR decomposition. Then

$$\mathcal{L}_{\text{align}}(B) = 1 - \tfrac{1}{r}\|Q_B^\top \widetilde{U}_{T,r}\|_F^2 = \tfrac{1}{r}\sum_{i=1}^{r} \sin^2\theta_i,$$

where θ_i are the principal angles between $\mathrm{span}(Q_B) = \mathrm{span}(B)$ and $\mathrm{span}(\widetilde{U}_{T,r})$. $\mathcal{L}_{\text{align}}(B) = 0$ iff every $\theta_i = 0$, iff $\mathrm{span}(B) = \mathrm{span}(\widetilde{U}_{T,r})$. By Step 2 of Appendix B.3, this is also equivalent to $\mathcal{S}(B) = 0$. □

Why $\mathcal{L}_{\text{align}} \neq \mathcal{S}$. The exact spectral form of the subspace miss is

$$\|(I - P_B)\widetilde{W}_T\|_F^2 \;=\; \sum_i \widetilde{\sigma}_i^2 \left(1 - \|P_B \widetilde{u}_i\|_2^2\right),$$

which is $\widetilde{\sigma}_i^2$-weighted. The alignment loss $\mathcal{L}_{\text{align}} = \frac{1}{r}\sum_{i=1}^{r} \sin^2\theta_i$ is unweighted (each angle contributes equally) and considers only the top r teacher directions. Thus $\mathcal{L}_{\text{align}}$ shares its zero set with $\mathcal{S}$ (Lemma B.1) but does not match its level sets. In practice this is desirable: $\mathcal{L}_{\text{align}}$ is differentiable in B, computable from a single thin QR plus one $r \times r$ matmul, and numerically stable, whereas a $\widetilde{\sigma}_i^2$-weighted version would require computing the full $\widetilde{u}_i$ basis at every step.

B.6. Coefficient Loss: Data-Weighted vs. Unweighted Forms

The principled within-subspace coefficient term implied by $\mathcal{C}(B, A) = \|P_B \widetilde{W}_T - BA\Sigma_x^{1/2}\|_F^2$ depends on the singular values of $BA\Sigma_x^{1/2}$, since $P_B \widetilde{W}_T$ has top-r singular values bounded by those of $\widetilde{W}_T$ projected onto $\mathrm{span}(B)$. The matching loss

$$\mathcal{L}_{\text{coeff}}^{\text{dw}}(B, A) \;=\; \frac{\left\|\sigma(R_B A \Sigma_x^{1/2}) - \widetilde{\sigma}_{T,r}\right\|_2^2}{\left\|\widetilde{\sigma}_{T,r}\right\|_2^2 + \varepsilon}$$

is therefore the data-weighted form. Computing it requires applying $\Sigma_x^{1/2}$, which is not stored; the cheapest equivalent is right-multiplication by the calibration sketch X_c (which has the same nonzero singular values up to the $\sqrt{n-1}$ factor), at cost $O(rnd_{\text{in}})$ per layer per step. We instead use

$$\mathcal{L}_{\text{coeff}}(B, A) \;=\; \frac{\left\|\sigma(R_B A) - \widetilde{\sigma}_{T,r}\right\|_2^2}{\left\|\widetilde{\sigma}_{T,r}\right\|_2^2 + \varepsilon},$$

which costs $O(r^2 d_{\text{in}})$ per layer per step (one $r \times d_{\text{in}}$ SVD) and equals $\mathcal{L}_{\text{coeff}}^{\text{dw}}$ when $\Sigma_x = I$. In the anisotropic case the two losses differ; both share the trivial zero $A = 0, \widetilde{\sigma}_{T,r} = 0$ but not their general level sets. Because our ablations identify $\mathcal{L}_{\text{coeff}}$ as auxiliary rather than load-bearing (Section 5.4), the practical cost of using the proxy is small. We confirmed in preliminary runs on SST-2, MRPC, and CoLA that swapping $\mathcal{L}_{\text{coeff}}$ for $\mathcal{L}_{\text{coeff}}^{\text{dw}}$ produces no consistent improvement.

C. Additional Experimental Details

C.1. Hyperparameters

Table 3 collects all SAD-LoRA hyperparameters. Optimization, calibration size, and energy threshold values are held fixed across all six GLUE tasks and all five ranks; only the LoRA rank itself varies in our rank-trajectory experiments.

Table 3. SAD-LoRA hyperparameters used throughout. Values are matched across baselines wherever they share a parameter (optimizer, LR schedule, KD temperature).

Parameter	Value	Notes
SAD-LoRA-specific		
Alignment weight α	0.2	Eq. 10
Coefficient weight β	0.05	Eq. 10; 0 in SAD-LoRA-Align
Distillation temperature τ	4.0	Eq. 9
KD weight λ_{KD}	1.0	Eq. 10
Calibration size n	1024	Phase 2 (Section 3)
Spectral energy threshold δ	0.01	Prop. 4.3
Stored ranks $r_{\max}$	64	Cap on per-layer SVD
Numerical ε	10^{-8}	Eq. 8
Optimizer (matched across all methods)		
Optimizer	AdamW	Decoupled weight decay
Weight decay	0.01	—
Learning rate (LoRA)	5×10^{-4}	—
Warmup fraction	6%	Linear, then linear decay
Batch size	32	SST-2, QNLI; 16 for small tasks
Max sequence length	128	Standard for GLUE
Mixed precision	fp16	Spectral ops in fp32
Seeds	$\{42, 123, 2024\}$	—

C.2. Optimization

We use AdamW with weight decay 0.01, learning rate 5×10^{-4} for LoRA parameters, linear warmup over the first 6% of steps, and linear decay thereafter. Mixed-precision training is enabled, with all spectral operations (QR, SVD, computation of $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{coeff}}$) executed in float32 for numerical stability.

C.3. Spectral Target Construction

For each adapted layer, Phase 1 computes $\Delta W_T^{(\ell)}$ from a reference RoBERTa-base teacher fine-tuned on the same task with the same data split. Phase 2 uses $n = 1024$ calibration samples drawn from the training set; activations are mean-centered before forming $Z^{(\ell)}$. Truncated SVD uses randomized SVD (Halko et al., 2011) when $d_{\text{out}} > 256$.

C.4. Per-layer Rank

We use a uniform rank r across layers in our main results, matching the convention of prior baselines and isolating the effect of the SAD-LoRA losses from rank-allocation choices. Proposition 4.3 suggests that allocating ranks per layer using the spectral tail of $\widetilde{W}_T^{(\ell)}$ (with $\delta = 0.01$) could yield further gains; preliminary results indicate small but positive effects on parameter efficiency, but a careful per-layer ablation is left to future work.

C.5. Subspace Alignment Across Tasks and Ranks

Table 4 reports the final post-training subspace alignment $A^{(\ell)} = \frac{1}{r}\|Q_B^{(\ell)\top}\widetilde{U}_{T,r}^{(\ell)}\|_F^2$, averaged over adapted layers and three seeds, for SAD-LoRA at each rank. Higher is better; 1.0 means $\operatorname{span}(B)$ exactly recovers the data-weighted teacher subspace. The pattern mirrors the GLUE results: tasks with the strongest task-specific spectral structure (RTE, MRPC, CoLA, STS-B) reach near-perfect alignment even at $r = 4$, while QNLI and SST-2 — which are nearly saturated by the pretrained backbone and therefore have flatter $\widetilde{W}_T$ spectra — require larger r to recover the same alignment.

Table 4. Mean post-training subspace alignment $\frac{1}{r}\|Q_B^\top \widetilde{U}_{T,r}\|_F^2$ for SAD-LoRA across tasks and ranks (averaged over adapted layers and three seeds). Bold entries indicate ≥ 0.95 alignment.

Task	r=1	r=2	r=4	r=8	r=16
RTE	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**
MRPC	0.87	0.90	0.94	**0.96**	**0.97**
CoLA	0.82	0.90	0.93	**0.95**	**0.97**
SST-2	0.75	0.84	0.88	0.91	0.94
QNLI	0.69	0.78	0.84	0.88	0.90
STS-B	0.68	0.82	0.90	0.94	**0.96**

C.6. Per-Task Ablation Across Ranks

Fig. 6 extends Fig. 5 by reporting all three SAD-LoRA variants — full SAD-LoRA, SAD-LoRA-Align (β=0), and SAD-LoRA-Coeff (α=0) — at both r=4 and r=8 for every GLUE task. The deltas are computed against the strongest non-SAD baseline (KD-LoRA, Logit-MSE, NN-Init, or PiSSA-Init) for each task and rank. Two patterns are visible: (i) SAD-LoRA-Coeff is the most damaging ablation, with large negative deltas on RTE at both ranks (-6.1 at r=4, -4.1 at r=8), confirming that subspace alignment carries the load; (ii) SAD-LoRA-Align is competitive across all tasks at r=8, while the full objective adds a measurable advantage on SST-2 (+0.5 at both ranks). MRPC is the consistent exception, where pretrained-weight principal structure (PiSSA-Init) remains a stronger inductive bias than the data-weighted teacher update.

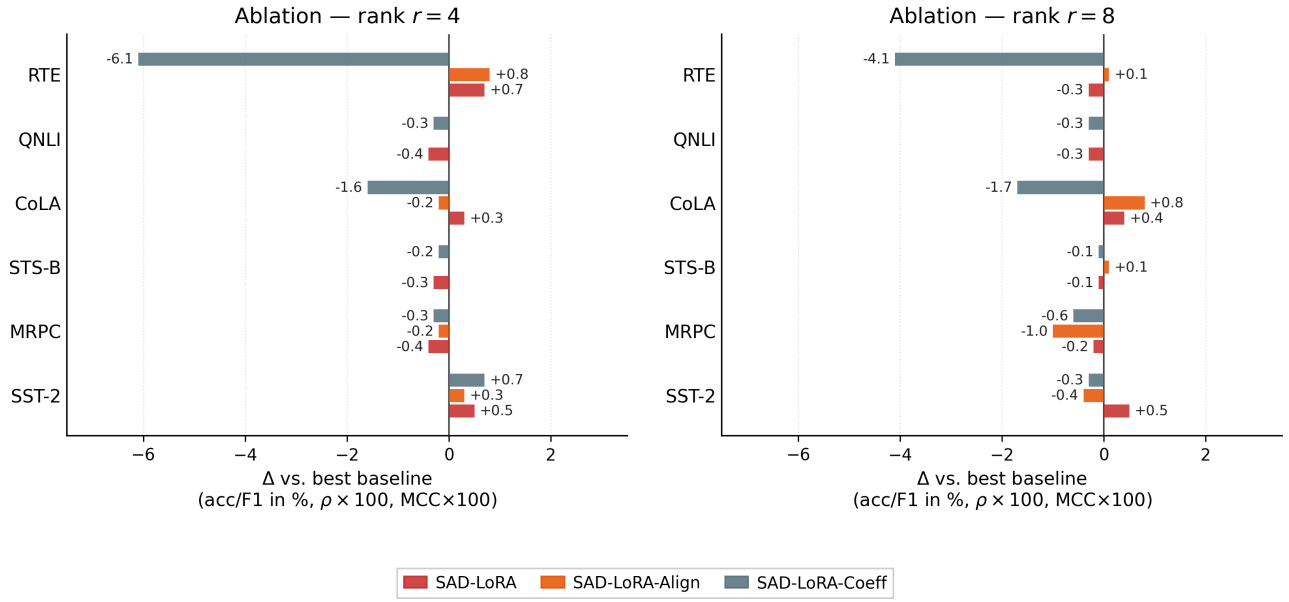

Figure 6. Per-task ablation: Δ versus the best non-SAD baseline at r=4 (left) and r=8 (right) for each SAD-LoRA variant. Positive values indicate improvement over the strongest baseline. Removing $\mathcal{L}_{\text{align}}$ (SAD-LoRA-Coeff) is consistently more damaging than removing $\mathcal{L}_{\text{coeff}}$ (SAD-LoRA-Align).

C.7. Reproducibility

Code is available at https://github.com/OmerTariq-KAIST/sad-lora. Random seeds are $\{42, 123, 2024\}$. Each GLUE configuration was trained on a single RTX6000Pro 98GB GPU; total wall-clock time for the GLUE result table is approximately 48 GPU-hours.