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

Omer Tariq Syed Muhammad Raza Jeongbae Son

ICML International Conference On Machine Learning

NEUBILITY

Core Motivation: Why Spectral Alignment?

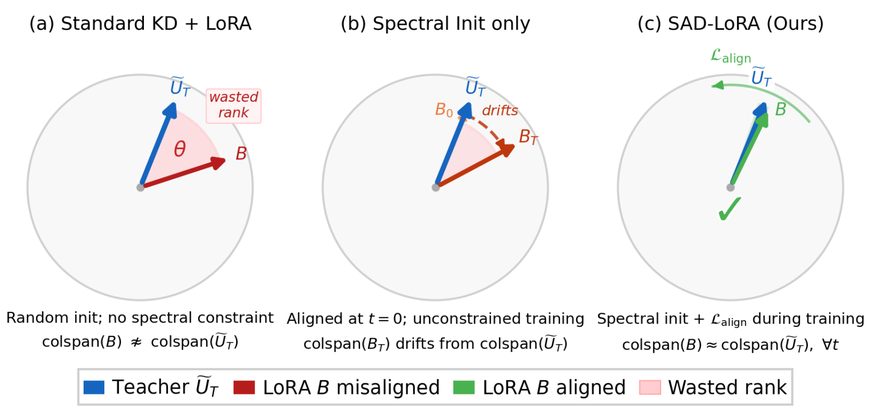
KD+LoRA compresses large models efficiently, but standard KD never tells the adapter which low-rank subspace to learn. As a result, valuable rank capacity is wasted.

SAD-LoRA explicitly **aligns the LoRA adapter** with the **teacher's data-weighted spectral subspace**, improving rank efficiency without additional inference cost.

✓ First framework that explicitly aligns the LoRA subspace during knowledge distillation

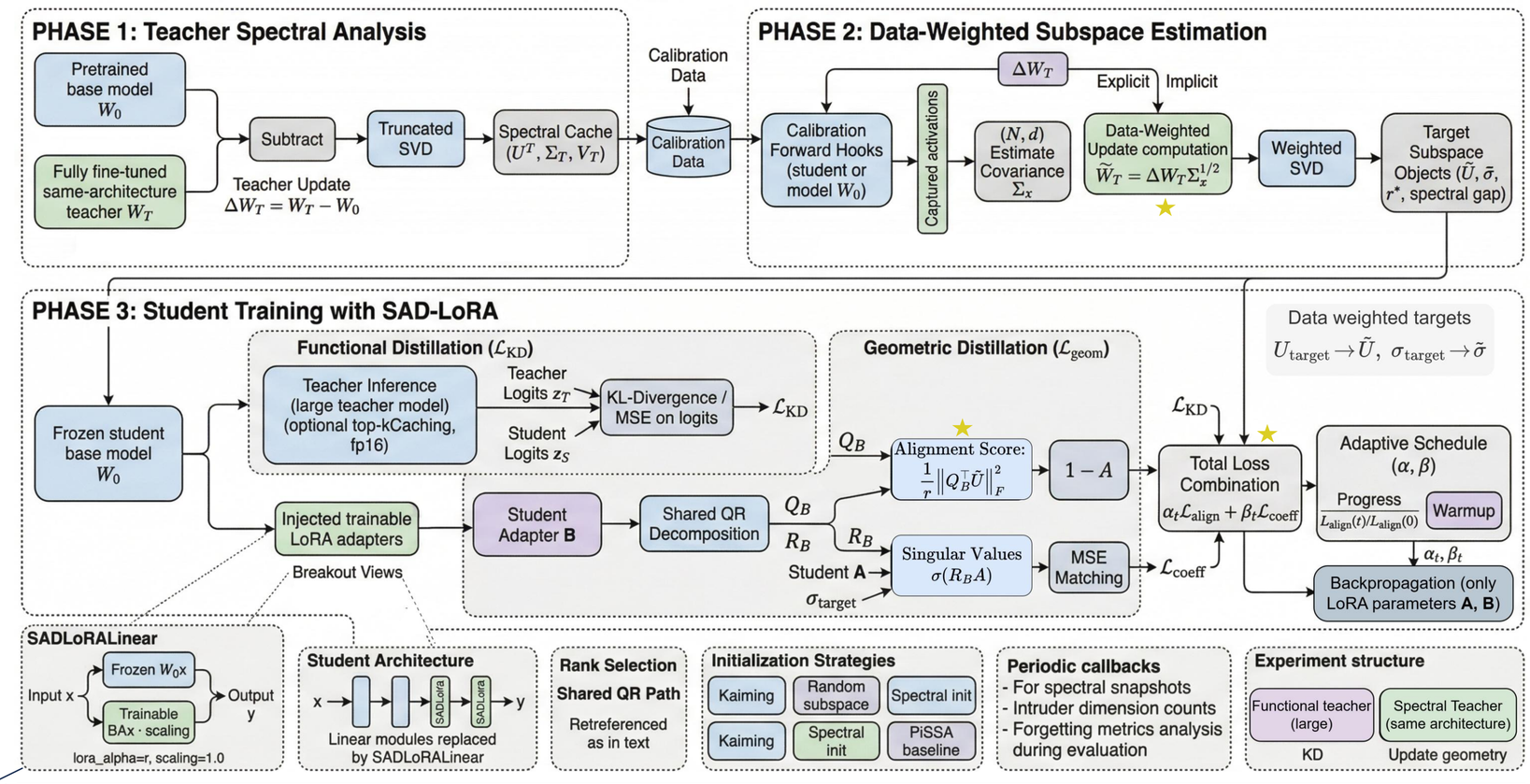
✓ Theoretically decomposes distillation error into subspace misalignment, coefficient mismatch, and irreducible rank residual

✓ Improves low-rank distillation across GLUE while introducing no additional inference overhead



Methodology: SAD-LoRA Pipeline

➡ Compute the teacher's data-weighted spectral subspace once, then continuously align the LoRA adapter to it throughout distillation



➡ Standard knowledge distillation learns outputs.

➡ **SAD-LoRA** additionally learns the optimal low-rank subspace, preventing subspace drift throughout training.

Theoretical Error Decomposition

Distillation Error Decomposes into Three Components

$$\mathcal{E} = \mathcal{E}_{\text{subspace}} + \mathcal{E}_{\text{coeff}} + \mathcal{E}_{\text{residual}}$$

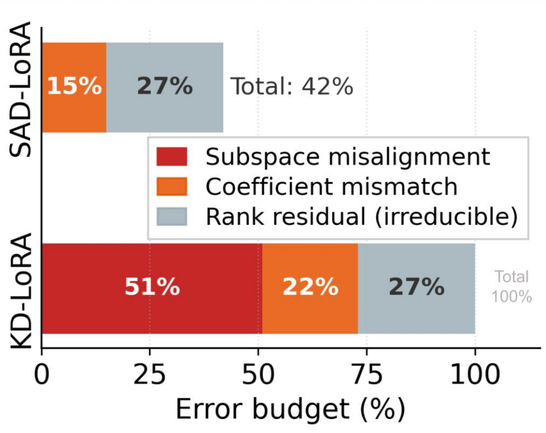
1. Subspace Misalignment (New)
Student learns the wrong directions (subspace error).
Minimized by $\mathcal{L}_{\text{align}}$

2. Coefficient Mismatch
Student uses correct directions but with incorrect strengths.
Minimized by $\mathcal{L}_{\text{coeff}}$

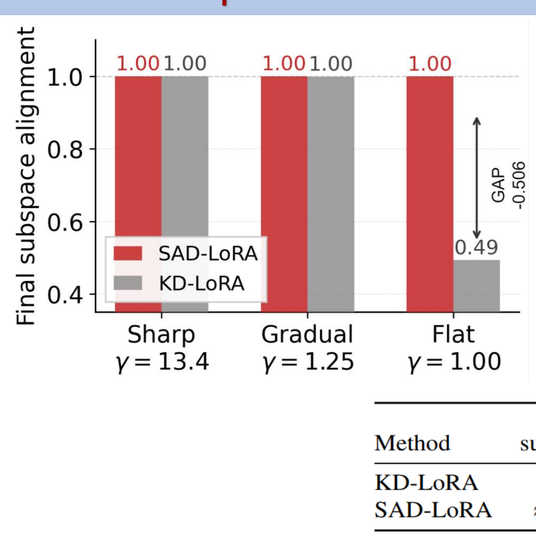
3. Rank Residual
Information that lies beyond the rank-r subspace.
Irreducible at rank r

Key Takeaways

- **Subspace misalignment** ($\mathcal{S}$) is the dominant source of error in KD-LoRA (51%).
- **SAD-LoRA** explicitly reduces subspace misalignment to ≈ 0%, lowering total error from 100% → 42%.
- **Rank residual** ($\mathcal{R}$) is irreducible at rank–r (27%) for both methods.

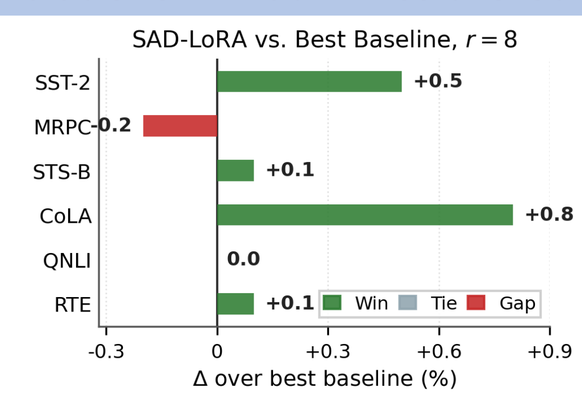


Controlled Spectral Validation



Method	subspace	coeff.	residual	Total
KD-LoRA	51%	22%	27%	100%
SAD-LoRA	≈ 0%	15%	27%	42%

Ablations: What Drives the Gain?



Sad-LoRA Training Objectives

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{KD}\mathcal{L}_{KD} + \frac{\alpha}{|S|}\sum_{\ell}\mathcal{L}_{\text{align}}^{(\ell)} + \frac{\beta}{|S|}\sum_{\ell}\mathcal{L}_{\text{coeff}}^{(\ell)}$$

$\mathcal{L}_{\text{align}}$ - principal-angle loss on $\text{span}(B)$ · *load-bearing*

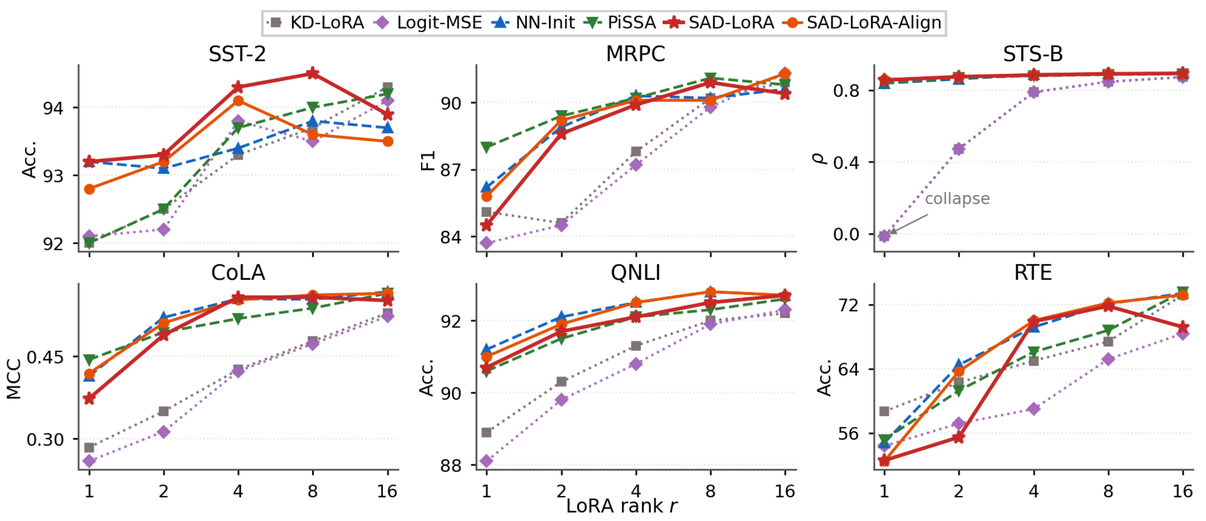
$\mathcal{L}_{\text{coeff}}$ - spectral coefficient matching · *auxiliary only*

Results

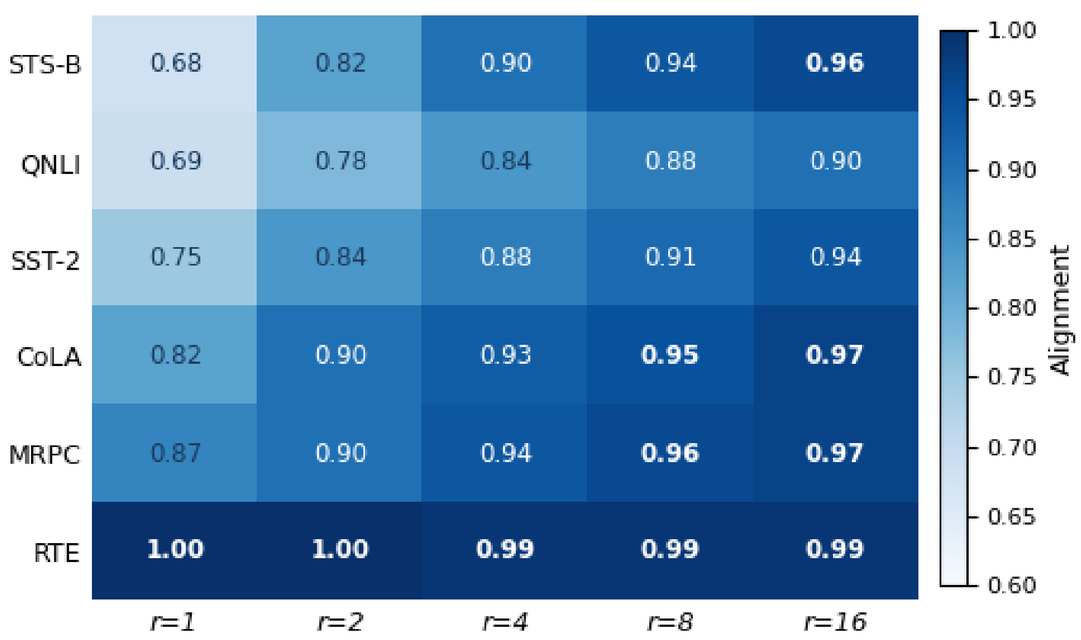
SAD-LoRA @ r=4 ≥ KD-LoRA @ r=8 on STS-B & CoLA — **better subspace beats more rank**

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

Rank Efficiency on GLUE



Post-Training Subspace Alignment



Limitations and Future Work

1. Requires same-arch reference teacher (One extra offline model)
Our approach relies on a teacher model with the same architecture, incurring the cost of training and maintaining one additional model offline.

2. Calibration pass: one-time, no inference overhead
A one-time calibration step is required to determine the optimal rank, but it does not introduce any inference-time overhead.

3. Future: decoder-only LLMs, generation tasks, per-layer rank
We plan to extend SAD-LoRA to decoder-only LLMs, evaluate on generation tasks, and explore adaptive, per-layer rank allocation.

Conclusion

1. **Subspace misalignment is the hidden bottleneck in KD+LoRA.** Standard KD cannot see or fix it — SAD-LoRA is the first method to explicitly target it via Grassmannian alignment loss.
2. **Better subspace beats more rank.** SAD-LoRA at $r = 4$ matches or exceeds KD-LoRA at $r = 8$ on STS-B and CoLA — subspace selection compensates for additional rank budget.
3. **$\mathcal{L}_{align}$ is load-bearing; $\mathcal{L}_{coeff}$ is auxiliary.** Ablations confirm: removing $\mathcal{L}_{align}$ costs up to ~6.1 on RTE (r=4). Removing $\mathcal{L}_{coeff}$ has negligible impact. Direction matters more than magnitude.
4. **Zero inference overhead — a drop-in for any KD+LoRA pipeline.** Spectral targets are computed once offline. No extra modules at inference. Adapts to any encoder-based model.

Sad-LoRA