PRETRAINING INDUCES STABLE SPECTRAL SUBSPACES FOR DOWNSTREAM TASK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Finetuning pretrained models often exhibits a low-dimensional structure in parameter space, with prior work focusing on where adaptation occurs. We instead investigate which parameter subspaces remain stable across finetuning and downstream tasks. We analyze pretrained weight matrices across vision and language models and find that leading spectral subspaces remain largely unchanged during finetuning, with stability consistently stronger in earlier layers. This stability emerges progressively during pretraining and increases with pretraining dataset size, suggesting that downstream adaptation largely preserves the leading spectral structure formed during pretraining. We further show that downstream adaptation can be restricted to pretrained leading spectral subspaces in deeper layers, achieving performance close to full finetuning while updating only 0.2% of model parameters on GLUE. Our findings offer a new parameter-space perspective on how pretraining shapes downstream adaptation, highlighting leading spectral subspaces as stable structure inherited across tasks.

1 INTRODUCTION

Finetuning a pretrained model on a downstream task is the main paradigm in modern machine learning (Wu et al., 2025; Han et al., 2024). A well-established observation is that parameter changes during finetuning are geometrically concentrated: despite updating all weights, finetuning often occurs in a much lower-dimensional subspace than the full parameter space (Aghajanyan et al., 2021; Zhang et al., 2023b; Fukuhata & Kano, 2025). This has motivated efforts to identify the directions that *change* during finetuning and to use them to design more parameter-efficient finetuning algorithms (Zhang et al., 2023b; Meng et al., 2024). Yet an equally important question remains less explored: what about the directions that *remain stable* during finetuning (Figure 1A)?

The directions that remain stable may be just as informative as those that change. If finetuning on different downstream tasks consistently preserves certain parameter directions, these directions may reflect information established during pretraining that is broadly useful across tasks. Identifying such stable directions and determining whether they are shared across downstream tasks could therefore provide insight into the parameter structure underlying the transferability of pretrained models.

We investigate this question using singular value decomposition (SVD), which provides a natural framework for characterizing both individual spectral directions and the subspaces they span (Figure 1B). We measure the alignment of corresponding singular vectors of parameter matrix before and after finetuning, and further characterize subspace similarity using principal angles, which remain well-defined when singular values are closely spaced.

Across diverse tasks, architectures, and modalities, we find a clear spectral and layer-wise pattern: leading singular directions and their corresponding subspaces are largely preserved after finetuning, particularly in earlier layers, whereas changes are more pronounced in deeper layers and lower-ranked directions (Figure 1C). Importantly, we find that the leading spectral subspaces are shared across downstream tasks: replacing the leading singular vectors of one task-specific model with those from another causes little change in downstream performance. This indicates that different tasks rely on a common dominant spectral structure. Thus, finetuning does not generally reorganize the entire parameter space. Instead, it preserves a shared spectral organization inherited from pretraining and modifies only a relatively small task-specific portion.

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.

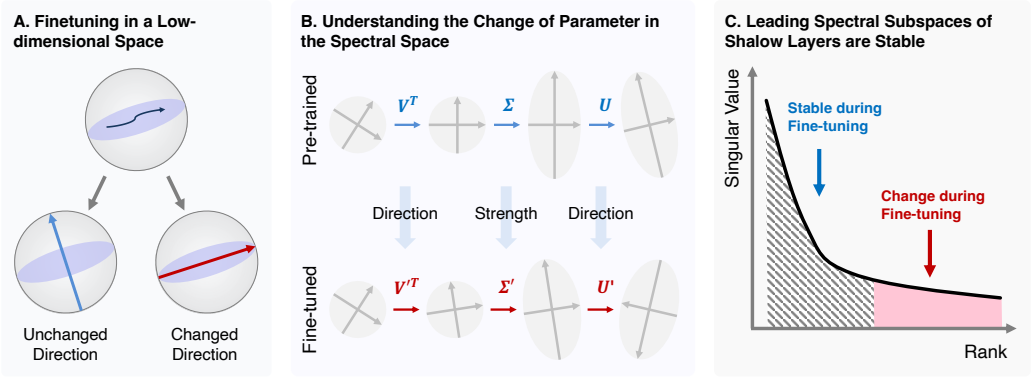


Figure 1: **Pretraining induces a stable spectral subspace for downstream task adaptation. (A)** Finetuning updates are not uniformly distributed across the parameter space: some directions undergo significant changes while others remain largely unchanged. **(B)** To characterize the spectral structure preserved during finetuning, we decompose weight matrices via SVD and examine changes at both the level of individual singular vectors and the subspaces spanned by singular vectors before and after finetuning. **(C)** Both leading singular vectors and their corresponding spectral subspaces exhibit substantial stability after finetuning, particularly in shallower layers.

We next investigate what influences the stability of this spectral structure. Since the scale of pretraining data is a key factor underlying the capabilities of pretrained models, we ask whether larger-scale pretraining also produces more stable parameter subspaces. Across both new downstream tasks and new data distributions, models pretrained on larger datasets consistently exhibit greater stability in their leading spectral subspaces. Moreover, these leading subspaces progressively stabilize throughout pretraining. These results suggest that the dominant spectral organization used by downstream models is gradually established during pretraining and becomes more stable as the pretraining data scale increases.

The observations above also motivate a functional test of how the identified spectral structure relates to downstream adaptation. Since earlier layers exhibit smaller changes while deeper layers undergo larger changes during finetuning, we first examine the contribution of different layers by adapting them separately. We find that downstream performance consistently improves as adaptation moves toward deeper layers, indicating that deeper layers contribute more to downstream adaptation. We then further restrict adaptation in these deeper layers to their leading spectral subspaces. Remarkably, this restricted adaptation achieves performance close to full finetuning across GLUE tasks (Wang et al., 2018), while updating only 0.2% of the model parameters, providing functional evidence that downstream adaptation can be effectively realized within the dominant spectral structure established during pretraining.

Our main contributions are as follows:

- We reveal a stable spectral structure in pretrained weights: leading singular directions and subspaces are largely preserved during finetuning, particularly in earlier layers, across diverse tasks, architectures, and modalities.
- We show that this structure progressively stabilizes during pretraining and becomes more stable under new tasks and data distributions as pretraining data scale increases.
- We provide functional evidence that effective finetuning can be achieved by adapting only leading spectral subspaces in deeper layers, validating the role of the dominant pretrained structure in downstream adaptation.

2 Related Work

Parameter-efficient fine-tuning (PEFT) and low-dimensional adaptation. Full-parameter fine-tuning often exhibits a low-dimensional structure, with parameter updates concentrated in a low-dimensional subspace of the full parameter space (Aghajanyan et al., 2021; Li et al., 2018; Zhang et al., 2023b; Fukuhata & Kano, 2025). This observation has motivated methods that restrict adaptation

to a small number of trainable parameters or low-dimensional update spaces, including adapters, LoRA, BitFit, and related approaches (Houlsby et al., 2019; Hu et al., 2022; Zaken et al., 2022; Zhang et al., 2023a; Liu et al., 2024; He et al., 2021; Meng et al., 2024). These studies primarily focus on the parameter directions that change during adaptation. We instead examine the complementary structure that remains stable, and ask whether such stability is shared across downstream tasks and shaped by pretraining.

Spectral analysis of neural network parameters. Spectral methods have been widely used to characterize the structure of neural network parameters. For example, prior work has shown that trained weight matrices often exhibit heavy-tailed singular-value spectra (Martin & Mahoney, 2021; Mahoney & Martin, 2019). While this line of work focuses primarily on describing the spectral properties of trained models, we use spectral analysis to study how parameter structure changes during finetuning.

Pretraining scale and transferability. Scaling-law studies have empirically established that larger pretraining datasets generally lead to better downstream performance (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023; Hernandez et al., 2021). Yet beyond its effect on performance, it remains unclear how increasing pretraining data shapes the underlying parameter structure that supports downstream transfer. We show that larger-scale pretraining produces more stable leading spectral subspaces, which can be preserved during downstream adaptation while restricting parameter changes to a small portion of the model. This provides a structural perspective on how increased pretraining data may contribute to the transferability of pretrained models.

3 Preliminaries

3.1 Spectral Structure and Similarity Measures

For a weight matrix $W \in \mathbb{R}^{m\times n}$, the SVD is $W = U\Sigma V^\top$, where the columns of U and V are the left and right singular vectors, and Σ contains the singular values in decreasing order. To characterize spectral changes after finetuning, we first measure the alignment between corresponding pretrained and finetuned singular vectors using absolute cosine similarity. Specifically, for the i-th component, we compute

$$a_i^{(u)} = \left|\tilde{u}_i^\top u_i\right|, \qquad a_i^{(v)} = \left|\tilde{v}_i^\top v_i\right|, \tag{1}$$

where u_i, v_i and $\tilde{u}_i, \tilde{v}_i$ denote the i-th left and right singular vectors of the pretrained and finetuned weight matrices, respectively. The absolute value removes the sign ambiguity inherent to singular vectors.

Because individual singular vectors can become unstable or non-identifiable when singular values are close, we further analyze the stability of the spectral structure at the subspace level. Specifically, we measure the similarity between the leading left and right singular subspaces of the pretrained and finetuned weight matrices using principal angles. Let $U_k, \tilde{U}_k \in \mathbb{R}^{m\times k}$ denote orthonormal bases for the top-k left singular subspaces of the pretrained and finetuned weight matrices, respectively. The cosines of the principal angles $\{\theta_j\}_{j=1}^k$ are given by

$$\cos\theta_j = \sigma_j\left(U_k^\top \tilde{U}_k\right). \tag{2}$$

We define the subspace similarity as

$$S_k^{(u)} = \frac{1}{k}\sum_{j=1}^{k}\cos\theta_j, \tag{3}$$

and analogously define $S_k^{(v)}$ for the right singular subspaces. Values close to 1 indicate highly similar subspaces. Unlike vector-level alignment, subspace similarity is invariant to rotations within the subspace and therefore provides a stable characterization of leading spectral structure even when individual singular vectors are not uniquely identifiable.

3.2 Spectrally Restricted Finetuning

Our spectral analysis reveals that finetuning induces larger changes in deeper layers, while the singular-value spectra of weight matrices decay rapidly, with a substantial fraction of their energy concentrated in the leading singular subspaces (Martin & Mahoney, 2021). These observations suggest that the leading spectral subspaces of deeper layers may provide an effective parameter space for downstream adaptation. We therefore restrict finetuning to these subspaces, allowing the model to adapt where parameter changes are most pronounced while retaining the dominant spectral structure of the pretrained weights.

Concretely, for each adaptable weight matrix, let $W^{\text{pt}} = U\Sigma V^\top$ be its pretrained SVD, and let U_R and V_R denote the leading R left and right singular vectors. We keep these pretrained singular vectors fixed and parameterize the finetuned weight as

$$W = W^{\text{pt}} + U_R M V_R^\top, \tag{4}$$

where $M \in \mathbb{R}^{R\times R}$ is trainable. The matrix M allows the model to reweight and mix the pretrained leading spectral components without introducing updates outside their span. This reduces the number of trainable parameters for an $m \times n$ weight matrix from mn to R^2, while retaining the dominant spectral structure identified by our analysis.

4 Experiments and Results

We organize our experiments around a central question: which spectral structures in pretrained weight matrices remain stable under finetuning, how this structure is formed and shared across downstream tasks, and whether it has functional implications for adaptation. We proceed in four steps. First, we establish the stability of leading spectral subspaces across diverse architectures and finetuning settings (**Section**4.1). Second, we examine whether the preserved spectral structure is shared across different downstream tasks (**Section** 4.2). Third, we trace how spectral stability develops during pretraining and how it depends on the scale of pretraining data (**Section** 4.3). Finally, we test whether the observed spectral organization can guide downstream adaptation by restricting parameter updates to selected pretrained spectral subspaces (**Section** 4.4).

4.1 Leading Spectral Subspaces Remain Stable During Finetuning

We first analyze how individual singular directions of pretrained weight matrices change after finetuning. Specifically, we measure the cosine similarity between corresponding pretrained and finetuned singular vectors across layers and spectral ranks. Across language models, singular-vector similarity remains consistently high across all layers. Vision models show similarly high similarity in earlier layers, but exhibit more pronounced changes in deeper layers (Figure 2A–B).

Figure 2C summarizes the cosine similarity averaged across layers over the evaluated spectral ranks. For vision models, the similarity remains above 0.8 across the analyzed ranks, while for language models it remains above 0.9. Figure 2D further shows the layer-wise evolution of the average similarity over the leading 30 singular vectors. In both vision and language models, similarity gradually decreases with depth, with a more pronounced decrease in the deeper layers of vision models. In contrast, language models maintain similarity above 0.9 even in the deeper layers. These results show that finetuning primarily alters individual spectral directions in deeper layers, while leaving them highly consistent across most of the network.

Because individual singular vectors can become unstable when singular values are closely spaced, we next repeat the analysis at the subspace level using principal-angle similarity, which provides a basis-invariant measure of spectral stability. Figures 2E–H show consistent patterns at the subspace level: the leading left and right singular subspaces remain highly similar before and after finetuning across the evaluated models and ranks. The layer-wise patterns further confirm that spectral changes are concentrated in deeper layers, particularly for vision models.

Together, the vector- and subspace-level analyses provide consistent evidence that finetuning largely preserves the dominant spectral structure established during pretraining, with changes concentrated in deeper layers and more pronounced in vision models. Details of the models and experimental settings

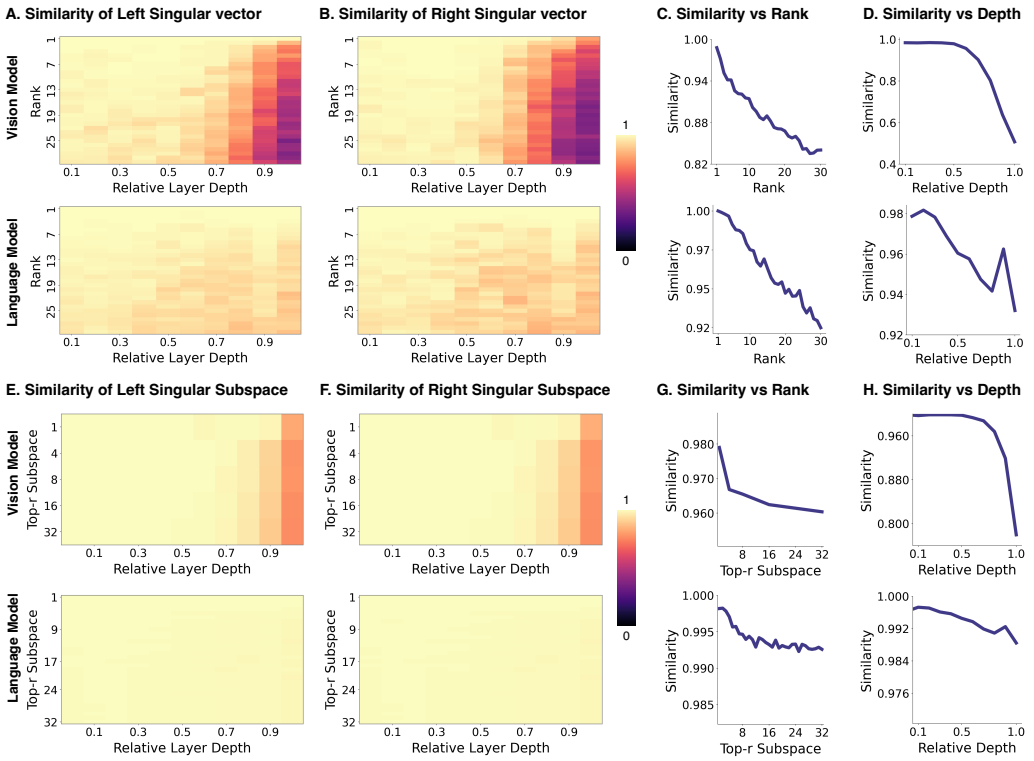


Figure 2: **Finetuning largely preserves the pretrained spectral subspaces. (A, B)** Cosine similarity of matched left and right singular vectors before and after finetuning across language and vision models. **(C, D)** Cosine similarity of singular vectors as a function of spectral rank and network depth respectively. **(E–H)** The corresponding analysis at the subspace level, showing principal-angle similarity between pretrained and finetuned spectral subspaces across rank and network depth.

are provided in Appendix A. Additional results with more spectral ranks are provided in Appendix D. We also analyze the corresponding singular values in Appendix B, finding that they remain largely unchanged before and after finetuning across most models. We further examine spectral changes in attention modules in Appendix F and Appendix G, with results consistent with those observed in FFN layers.

4.2 Shared Spectral Structure Across Downstream Tasks

The stability of pretrained spectral structure raises a further question: whether the preserved structure is shared across different downstream tasks. We first examine whether the leading spectral components of a finetuned model can be replaced by those of its pretrained counterpart without substantially affecting task performance. We conduct this experiment on `bert-base-uncased` finetuned on four downstream tasks: COLA, MNLI, QNLI, and SST2. Specifically, we replace the top-r singular vectors of the finetuned parameters with the corresponding singular vectors from the pretrained model while keeping the finetuned singular values unchanged. As shown in Figure 3A, restoring the pretrained singular vectors causes only minor changes in performance across all four tasks. In contrast, replacing these singular vectors with random vectors or zero vectors leads to substantial performance degradation. These results indicate that the leading spectral structure preserved during finetuning can be restored from the pretrained model without substantially disrupting downstream functionality.

We next test whether this preserved spectral structure is shared across downstream tasks. For each of the four tasks, we replace the top-r singular vectors of its finetuned model with the corresponding singular vectors from models finetuned on each of the other three tasks. As shown in Figure 3B, these cross-task replacements lead to only small changes in performance relative to the original finetuned

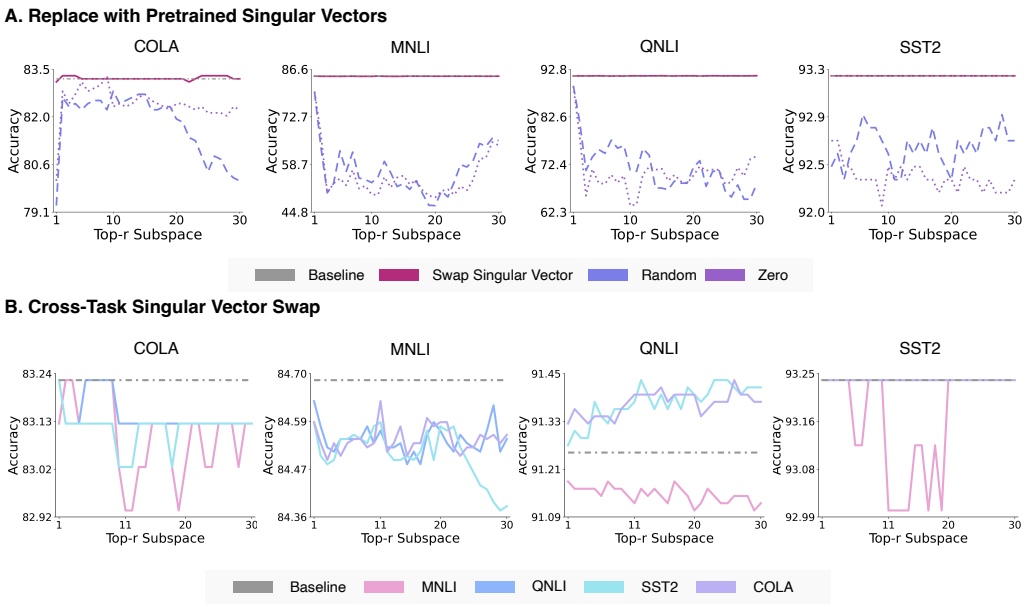


Figure 3: **Finetuning largely preserves the leading pretrained spectral subspaces. (A)** Replacing finetuned spectral components with their pretrained counterparts causes little performance loss, whereas random replacement or zeroing leads to substantial degradation. **(B)** Swapping leading spectral components across models finetuned on different downstream tasks has limited impact on performance.

models. Although the replacement can produce measurable changes, the overall performance remains largely unchanged across tasks. The limited effect of cross-task replacement suggests that models finetuned on different tasks retain a substantial portion of a common spectral structure inherited from pretraining.

Together, these replacement experiments provide functional evidence that the stable spectral structure identified above is shared across downstream tasks. This suggests that pretraining establishes a transferable spectral structure that is largely retained across downstream adaptation, while task-specific learning can occur without substantially reorganizing the dominant spectral subspaces. The main analysis focuses on FFN layers, corresponding results for attention modules are provided in Appendix H, showing consistent patterns. We further examine the functional importance of different spectral ranks in Appendix J by progressively masking singular vectors from either the head or the tail of the spectrum. Masking from the tail has little effect until a large number of components are removed, whereas masking from the head causes performance degradation after removing only a small number of leading components, further highlighting the functional importance of the leading spectral structure.

4.3 Larger Pretraining Datasets Produce More Stable Spectral Subspaces

The previous sections show that finetuning largely preserves the leading spectral subspaces of pretrained models and that these preserved components are shared across downstream tasks. We next examine how this stable spectral structure develops during pretraining and whether its stability depends on the amount of pretraining data.

We first track the stability of individual singular vectors during pretraining. We train a Vision Transformer on a masked image reconstruction task for 50 epochs and extract the top singular vectors of the FFN weights from the final layer at each checkpoint. We then compute pairwise cosine similarities between singular vectors across training epochs. As shown in Figure 4 A, the singular vectors become increasingly similar as pretraining proceeds, with the final 10 epochs exhibiting consistently high pairwise similarity. This indicates that the dominant spectral directions progressively stabilize during pretraining.

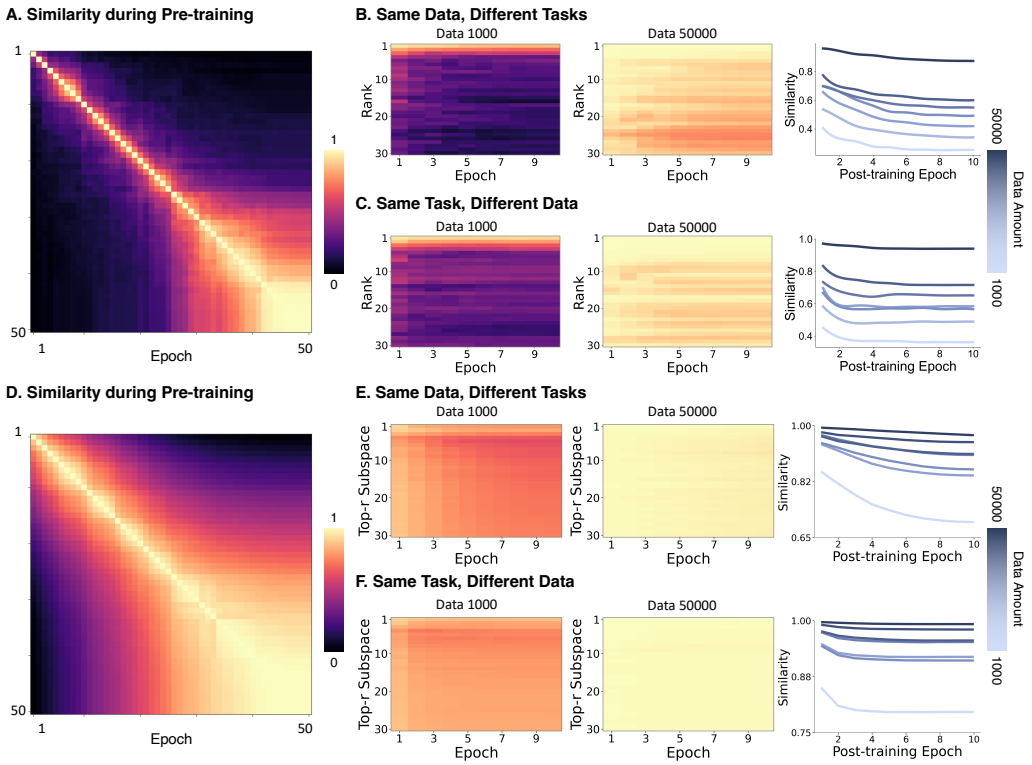


Figure 4: **Pretraining progressively stabilizes the spectral structure, and larger datasets improve its transferability.** **(A)** Similarity between leading spectral components from different pretraining checkpoints. Similarity increases in later stages, showing that the dominant spectral structure progressively stabilizes during pretraining. **(B)** Post-training on different tasks with the same data source. **(C)** Post-training on different data with the same task. In both settings, models pretrained on larger datasets exhibit smaller changes in spectral structure, indicating greater stability. **(D–F)** The corresponding analyses at the subspace level, which yield consistent results with the component-level analyses.

We next examine whether the scale of pretraining data affects this stability under post-training adaptation. Using the same architecture and training procedure with different amounts of pretraining data, we consider two settings: classification on the same data source, which changes the task (Figure 4 B), and masked image reconstruction on a new dataset, which changes the data distribution while keeping the task fixed (Figure 4 C). Across spectral ranks, models pretrained on larger datasets consistently maintain higher singular-vector similarity as post-training proceeds in both settings. Thus, larger-scale pretraining produces spectral directions that undergo smaller changes under both task and data shifts.

We then repeat the same analyses at the subspace level, defining each spectral subspace as the span of the top 32 singular vectors and measuring similarity using principal angles. Figures 4 D–F show consistent results: these leading spectral subspaces progressively stabilize during pretraining, and larger pretraining datasets lead to higher subspace similarity under both changes in task and changes in data distribution. Together, these results show that increasing pretraining data is associated with a progressively more stable spectral structure that is preserved under subsequent adaptation. More results and details are provided in Appendix K.

4.4 Pre-trained Spectral Subspaces Support Efficient Transfer

Our analyses above reveal a consistent structure in how pretrained spectral subspaces are preserved and modified during finetuning. Leading spectral subspaces remain highly stable, particularly in earlier layers, whereas spectral changes become more pronounced in deeper layers. Meanwhile,

pretrained weight matrices typically exhibit rapidly decaying singular values, with substantial spectral energy concentrated in the leading components (Martin & Mahoney, 2021). These observations suggest that downstream adaptation may be largely supported by a relatively small set of pretrained spectral coordinates, particularly in deeper layers.

Here, we use spectrally restricted finetuning to test this implication. Our goal is not to introduce a new PEFT method or establish a new state-of-the-art result, but to provide an engineering-level validation of the spectral structure identified above. Specifically, we ask whether substantial downstream adaptation can be achieved when learning is confined to the pretrained leading spectral subspaces, thereby testing whether the spectral structure observed under full finetuning has practical value for transfer.

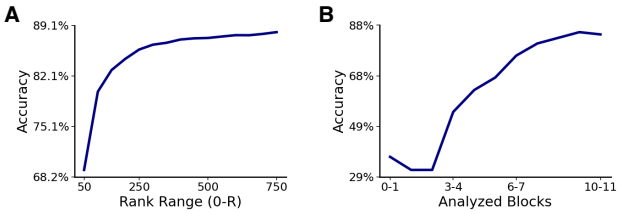


Figure 5: **Spectral subspace capacity and adaptation location affect finetuning performance.** **(A)** Increasing the dimension of the adapted spectral subspace improves downstream performance before reaching saturation. **(B)** Allocating the adaptation budget to deeper layers leads to better performance.

We first examine how much spectral capacity is required for adaptation. As shown in Figure 5A, performance increases with K and largely saturates around $K = 300$, indicating that a relatively compact subset of the leading spectral coordinates is sufficient for substantial adaptation.

We next examine whether the layer dependence of spectral changes is reflected in the location of useful adaptation. Fixing the adaptation rank, we separately adapt each transformer block while keeping the remaining layers fixed. As shown in Figure 5B, adaptation becomes more effective in deeper blocks, consistent with our earlier observation that deeper layers undergo larger spectral changes during finetuning.

Table 1: Comparison with representative PEFT methods on the GLUE benchmark using DeBERTa-v3-base. LoRA-K and LoRA-G denote LoRA with Kaiming and Gaussian initialization for B respectively.

Method	Params	RTE	QQP	CoLA	MRPC	SST-2	QNLI	MNLI	STS-B	Mean
Full FT	184M	83.75	92.40	69.19	89.46	95.63	94.03	89.90	91.60	88.25
PiSSA	1.33M	88.69	92.33	73.12	91.50	96.22	94.43	90.37	92.00	89.83
LoRA-K	1.33M	84.84	92.03	70.69	90.28	95.64	93.84	89.96	91.68	88.62
LoRA-G	1.33M	85.20	91.99	69.82	89.95	94.95	93.87	90.65	91.60	88.50
AdaLoRA	1.27M	88.09	92.23	71.45	90.69	96.10	94.55	90.76	91.84	89.46
DoRA	1.27M	86.04	92.07	70.85	90.93	95.79	94.10	90.29	91.79	88.98
HAdapter	1.22M	84.48	91.91	68.64	89.95	95.53	94.11	90.13	91.48	88.28
PAdapter	1.18M	85.20	92.04	68.77	89.46	95.61	94.29	90.33	91.54	88.41
SVFT-Band	0.36M	81.26	84.53	62.97	82.35	94.84	91.83	87.22	85.46	83.81
BitFit	0.1M	78.70	88.41	66.96	87.75	94.84	92.24	89.37	91.35	86.20
SVFT-Plain	0.003M	62.45	85.21	49.79	75.73	90.82	80.81	78.33	81.40	75.57
Ours	0.36M	84.12	89.22	68.13	88.48	95.53	92.06	87.45	88.85	86.73

Finally, we instantiate these observations in a practical downstream setting. We adapt the top $K = 300$ spectral components in the FFN modules of the final two transformer blocks and evaluate this setting on the GLUE benchmark using DeBERTa-v3-base. With only 0.36M trainable parameters, corresponding to approximately 0.2% of the full model, this setting achieves an average score of 86.73, compared with 88.25 for full finetuning (Table 1). Representative PEFT methods are included as contextual benchmarks rather than as competing methods.

Together, these experiments provide a functional validation of the spectral organization identified in the preceding sections. The rank and layer dependence observed under full finetuning translate into a corresponding pattern of efficient downstream adaptation: a relatively small leading spectral subspace in deeper layers can support substantial task-specific learning without substantially changing the underlying pretrained spectral structure. More details are provided in Appendix L.

5 Discussion

Our results across vision and language models show that pretraining establishes stable spectral structures in the parameter space that are largely preserved during downstream finetuning and transferable across different tasks. This stability is consistently observed at both the singular-vector and subspace levels, suggesting that downstream adaptation often reuses rather than substantially reorganizes the dominant parameter subspaces formed during pretraining. The cross-task swapping experiments further demonstrate that these spectral components can support different downstream objectives, providing a possible explanation for why a single pretrained model can adapt effectively to diverse tasks. From this perspective, our findings highlight an important structural property underlying PEFT: effective adaptation may rely on a small set of pretrained spectral directions that already encode information broadly useful across downstream tasks, rather than requiring substantial reorganization of the full parameter space.

Our experiments further suggest that the stability of these spectral subspaces is related to pretraining data scale. Increasing the pretraining dataset size leads to more stable leading spectral subspaces, indicating that larger pretraining datasets establish parameter structures that are more consistently transferable across downstream tasks. The randomized-label experiments in Appendix E further show that this stability depends on meaningful task structure: when such structure is removed, finetuning induces substantially larger changes in the pretrained spectral subspaces. Together, these results suggest that spectral stability reflects the reuse of pretrained information during adaptation, linking pretraining data scale to spectral stability and cross-task transferability.

Limitations. Despite these observations, the current study remains primarily empirical and does not yet provide a theoretical explanation for the emergence of spectral stability. In particular, it remains unclear why larger pretraining datasets lead to more stable and transferable spectral subspaces, or why stability is concentrated in the leading spectral subspaces rather than being distributed uniformly across the spectrum. While our experiments provide empirical evidence for these phenomena, the underlying relationship between pretraining, data scale, spectral organization, and downstream adaptation remains to be formally characterized.

Future directions. Our findings suggest two directions for future work. First, the observed stability of pretrained spectral subspaces may provide a structural basis for designing more efficient PEFT methods. Rather than treating the update space as an arbitrary low-dimensional parameterization, future methods could explicitly exploit the stable spectral structure established during pretraining to identify and adapt only the most transferable directions. Second, our results raise the possibility of designing pretraining objectives, data mixtures, and model architectures that explicitly encourage the emergence of stable spectral subspaces that generalize across downstream tasks. Understanding and controlling such structures during pretraining may provide a new route toward improving the transferability and generalization of pretrained models.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7319–7328, 2021.

Shion Fukuhata and Yoshinobu Kano. Few dimensions are enough: Fine-tuning bert with selected dimensions revealed its redundant nature. *arXiv preprint arXiv:2504.04966*, 2025.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Seulgi Jeong, Yunseong Cho, and Sanghun Park. H-adapter: Pose-robust hairstyle transfer via attention-derived, source-aligned hair masks. In *European Conference on Computer Vision*, pp. 575–592. Springer, 2026.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham K Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. Svft: Parameter-efficient fine-tuning with singular vectors. *Advances in neural information processing systems*, 37:41425–41446, 2024.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pp. 4284–4293. PMLR, 2019.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37: 121038–121072, 2024.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Benjamin Newman, Prafulla Kumar Choubey, and Nazneen Rajani. P-adapters: Robustly extracting factual information from language models with diverse prompts. *arXiv preprint arXiv:2110.07280*, 2021.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 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.

Xiao-Kun Wu, Min Chen, Wanyi Li, Rui Wang, Limeng Lu, Jia Liu, Kai Hwang, Yixue Hao, Yanru Pan, Qingguo Meng, et al. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing*, 9(4):87, 2025.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 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 (Volume 2: Short Papers)*, pp. 1–9, 2022.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023a.

Zhong Zhang, Bang Liu, and Junming Shao. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1701–1713, 2023b.

AI USE DISCLOSURE

In this work, we used AI tools to improve the readability and clarity of the manuscript, including proofreading, grammatical correction, and refinement of phrasing. We reviewed and verified all AI-assisted text and take full responsibility for the final content, claims, and conclusions of this work.

APPENDIX OVERVIEW

This appendix provides additional empirical results, model details, and implementation specifications that further support the main claim of the paper: finetuning largely operates within a stable spectral subspace established by pretraining.

The supplementary material is organized around five broad questions.

First, **how general is spectral stability across model families, modalities, parameter types, and pretraining scales?** We first summarize the pretrained vision and language models used throughout our large-scale analyses and provide their downstream evaluation details. We then extend the spectral analysis beyond the FFN layers and standard vision and language settings. In particular, we examine spectral gaps across all FC layers in representative vision and language models, analyze attention projections in both modalities, and study the effect of pretraining data scale on spectral stability. These results show that the observed spectral structure is not restricted to a particular architecture, parameter type, or pretraining scale.

Second, **does spectral stability persist beyond the leading singular vectors, and how should it be characterized when individual vectors become ambiguous?** We therefore extend the rank-wise analysis from the leading components to the top-200 singular vectors and examine both individual singular-vector alignment and subspace similarity. We additionally analyze the spectral gaps of FC weight matrices, showing that neighboring singular values become closely spaced at relatively low ranks. This motivates the use of principal-angle similarity as a complementary, basis-invariant measure of spectral stability. Across both vector- and subspace-level analyses, the leading spectral components remain consistently more stable than lower-ranked components.

Third, **how do the singular values and different spectral ranks contribute to downstream adaptation?** Beyond singular-vector alignment, we analyze the singular-value spectra before and after finetuning and quantify their relative changes using multiple spectral statistics. For most pretrained models, finetuning introduces only modest changes in the singular-value distribution, although stronger rescaling can occur in models with larger pretraining–finetuning shifts. We further quantify the variance explained by different spectral ranks and perform masking experiments that compare removing leading versus trailing singular vectors. These results show that leading spectral components have a disproportionately large effect on downstream task performance despite accounting for only a minority of the total spectral energy.

Fourth, **when does spectral stability break down, and what does this reveal about the relationship between pretraining and downstream objectives?** We examine a controlled failure case using random-label finetuning, where the input data and optimization procedure are preserved but meaningful label structure is removed. Compared with standard finetuning, random-label adaptation produces substantially larger changes in both individual singular vectors and leading spectral subspaces. This result indicates that spectral stability is not simply a consequence of initialization or continued optimization, but is associated with the structure shared between pretraining and downstream learning.

Finally, **does the observed spectral stability have functional consequences for parameter-efficient transfer?** We provide additional singular-vector swapping experiments in attention modules and examine whether pretrained spectral directions can be directly reused across finetuned models and downstream tasks. We further provide implementation details and broader benchmark results for a spectrally restricted PEFT method, in which only a small number of coordinates in the pretrained singular-vector basis are optimized. Together, these experiments show that downstream adaptation can be effectively supported by a compact subset of pretrained spectral coordinates, providing a functional complement to the observed spectral stability.

Taken together, the appendix provides evidence from complementary perspectives: spectral-vector alignment, subspace stability, spectral-gap analysis, singular-value preservation, rank-wise variance, masking, controlled failure cases, cross-task swapping, pretraining data scaling, and spectrally restricted parameter-efficient transfer. These results collectively strengthen the central interpretation of the paper: pretraining establishes a transferable spectral coordinate system, and downstream finetuning largely adapts within this coordinate system by modifying a relatively small set of its dominant spectral components rather than replacing the underlying basis.

Roadmap of the appendix.

- **Section A: Details of Pretrained Models.** Summary of the pretrained vision and language models used in the large-scale analyses, including model families, pretraining sources, downstream performance, and model configurations.
- **Section B: Results of Singular Values.** Complementary analysis of singular-value spectra before and after finetuning, quantifying the extent to which finetuning preserves or rescales the pretrained spectral magnitudes.
- **Section C: Spectral Gap Analysis.** Analysis of relative spectral gaps across FC weight matrices from all layers of representative vision and language models, motivating the complementary use of subspace-level similarity when individual singular vectors become less identifiable.
- **Section D: Spectral Stability over Broader Rank Ranges.** Extended rank-wise analysis of singular-vector and spectral-subspace similarity across a broader range of ranks, showing that leading components remain more stable than lower-ranked directions.
- **Section E: Failure Cases of Spectral Stability.** Controlled random-label experiments showing that spectral stability substantially weakens when meaningful downstream task structure is removed.
- **Section F: Singular-Vector Analysis in Attention Modules.** Extension of the main FFN-layer analysis to query, key, value, and output projection matrices in attention modules across vision and language models.
- **Section G: Subspace Analysis in Attention Modules.** Principal-angle analysis of the leading spectral subspaces of attention projections, providing a basis-invariant complement to the singular-vector analysis.
- **Section H: Additional Results on Singular-Vector Swapping.** Extended task-level results for singular-vector replacement and cross-task swapping in attention modules, complementing the FFN results presented in the main text.
- **Section I: Information Content Captured by Top-K Singular Vectors.** Analysis of the cumulative explained variance associated with different spectral ranks, contextualizing how much of the total parameter energy is represented by the leading components.
- **Section J: Sensitivity of Masking Order.** Comparison of top-down and bottom-up singular-vector masking, showing that removing leading components has a substantially larger effect on downstream performance than removing trailing components.
- **Section K: Additional Results on the Effect of Pretraining Data Scale on Spectral Stability.** Expanded experiments across multiple pretraining data scales, showing that larger pretraining datasets are associated with greater singular-vector and subspace stability during subsequent adaptation.
- **Section L: Implementation Details for Spectral Parameter-Efficient Transfer.** Detailed implementation specifications for the spectrally restricted PEFT experiments, including spectral parameterization, target layers, training configurations, and GLUE benchmark settings.

A Details of Pretrained Models

This appendix summarizes the pretrained models used in our analyses. The main text focuses on vision and language models, demonstrating spectral transfer across diverse architectures, parameter scales, and pretraining regimes. In this appendix, we provide complete model specifications for both the vision and language models analyzed in the main text.

For vision, we study ConvNeXt models spanning multiple model sizes and pretraining sources, including ImageNet-12K, ImageNet-22K, and CLIP-style LAION pretraining. For language, we

Table 2: Top-1 accuracy of large-scale pretrained vision models on ImageNet-1K.

Pretrained Model	Finetuned Model Name	Acc (%)
convnext_nano.in12k	convnext_nano.in12k_ft_in1k	87.86
convnext_tiny.in12k	convnext_tiny.in12k_ft_in1k	88.91
convnext_tiny.fb_in22k	convnext_tiny.fb_in22k_ft_in1k	88.02
convnext_small.fb_in22k	convnext_small.fb_in22k_ft_in1k	89.93
convnext_small.in12k	convnext_small.in12k_ft_in1k	89.03
convnext_base.clip_laion2b_augreg	convnext_base.clip_laion2b_augreg_ft_in1k	89.68
convnext_large_mlp.clip_laion2b_augreg	convnext_large_mlp.clip_laion2b_augreg_ft_in1k	90.35
convnext_large.fb_in22k	convnext_large.fb_in22k_ft_in1k	90.31
convnext_xlarge.fb_in22k	convnext_xlarge.fb_in22k_ft_in1k	90.10
convnext_xxlarge.clip_laion2b_soup	convnext_xxlarge.clip_laion2b_soup_ft_in1k	90.67
timm/vit_base_patch16_224.augreg_in21k	timm/vit_base_patch16_224.augreg_in21k_ft_in1k	84.50
timm/vit_large_patch16_224.augreg_in21k	timm/vit_large_patch16_224.augreg_in21k_ft_in1k	85.90
timm/swin_base_patch4_window7_224.ms_in22k	timm/swin_base_patch4_window7_224.ms_in22k_ft_in1k	85.20
timm/swin_large_patch4_window7_224.ms_in22k	timm/swin_large_patch4_window7_224.ms_in22k_ft_in1k	86.30

Table 3: Accuracy of pretrained language models fine-tuned on the MNLI benchmark.

Pretrained Model	Finetuned Model Name	Acc (%)
google-bert/bert-base-uncased	ishan/bert-base-uncased-mnli	84.60
google-bert/bert-large-uncased	madlag/bert-large-uncased-mnli	86.05
google-bert/bert-base-cased	gchhablani/bert-base-cased-finetuned-mnli	84.60
google-bert/bert-large-cased	George-Ogden/bert-large-cased-finetuned-mnli	86.09
FacebookAI/roberta-base	textattack/roberta-base-MNLI	87.60
FacebookAI/roberta-large	FacebookAI/roberta-large-mnli	90.20
xlnet-base-cased	vish88/xlnet-base-mnli-finetuned	86.80
xlnet-large-cased	ynie/xlnet-large-cased-snli_mnli_fever_anli_R1_R2_R3-nli	90.80
microsoft/deberta-large	microsoft/deberta-large-mnli	91.30
microsoft/deberta-xlarge	microsoft/deberta-xlarge-mnli	91.50
microsoft/deberta-v2-xlarge	microsoft/deberta-xlarge-v2-mnli	91.70
microsoft/deberta-v2-xxlarge	microsoft/deberta-xxlarge-v2-mnli	91.90
microsoft/deberta-v3-base	MoritzLaurer/DeBERTa-v3-base-mnli	90.60
microsoft/deberta-v3-large	khalidalt/DeBERTa-v3-large-mnli	91.80
distilbert/distilbert-base-uncased	typeform/distilbert-base-uncased-mnli	88.80
google/bigbird-roberta-base	l-yohai/bigbird-roberta-base-mnli	87.50
google/mobilebert-uncased	Alireza1044/mobilebert_mnli	84.00
google/electra-base-discriminator	howey/electra-base-mnli	82.20

study a diverse collection of pretrained encoders and encoder-decoder variants, ranging from compact BERT-style models to large DeBERTa models. Together, these models allow us to test whether the spectral stability observed in the main text persists across substantial variation in architecture, scale, and pretraining data.

Table 2 lists the vision models and their Top-1 accuracy on ImageNet-1K. These models cover both standard supervised pretraining and large-scale multimodal pretraining, and are used in our cross-scale spectral analyses in vision.

Table 3 lists the language models and their downstream accuracy on MNLI. The collection includes BERT, Roberta, Xlnet and DeBERTa variants, providing a broad range of language-model scales and pretraining recipes for our transfer analysis.

B Results of Singular Values

The main text focuses on the stability of singular vectors during finetuning. Here we provide a complementary analysis of singular values. For each pretrained weight matrix and its finetuned counterpart, we compute the singular values and compare their changes after finetuning. Overall, we find that the singular-value spectra are also highly stable for most models: finetuning typically introduces only mild rescaling of the pretrained spectrum, rather than substantially reshaping it.

Figure 6 shows the relative change in singular values across different ranks for models in our analysis. For each model and rank, we compute the relative change as the difference between finetuned and

pretrained singular values divided by the pretrained value, visualized along the quantile axis. In most cases, the relative changes remain small across all ranks, indicating that finetuning preserves both the leading singular vectors and their spectral magnitudes. Noticeable deviations appear only in a small subset of models. Notably, these exceptions are primarily vision-language multimodal models, where adaptation may require stronger rescaling of pretrained features.

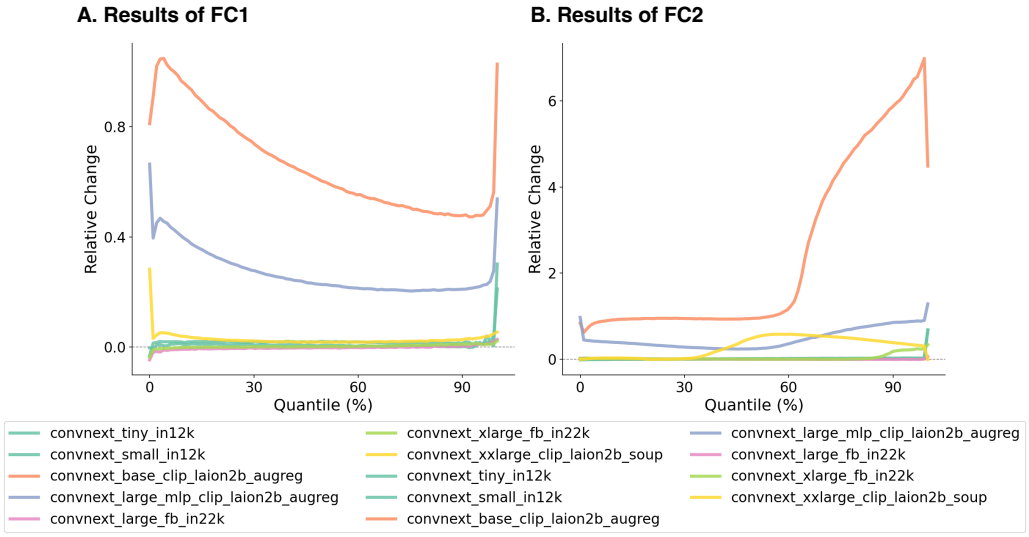


Figure 6: **Singular-value spectra before and after finetuning.** The pretrained and finetuned spectra are highly similar for most models, while larger spectral rescaling mainly appears in models with stronger pretraining–finetuning distribution shifts, such as multimodal pretrained models adapted to ImageNet.

To quantify these observations, Table 4 reports spectral change statistics for the last-block MLP weights of models at different scales. We use the relative spectral distance, $\mathrm{RSD} = \frac{\|\mathbf{s}_{\mathrm{ft}} - \mathbf{s}_{\mathrm{pre}}\|_2}{\|\mathbf{s}_{\mathrm{pre}}\|_2}$, to measure the overall change in the singular-value vector, where $\mathbf{s}_{\mathrm{pre}}$ and $\mathbf{s}_{\mathrm{ft}}$ denote the pretrained and finetuned singular-value vectors, respectively.

We further report the mean relative change (MRC) and maximum relative change (MaxRC) to summarize elementwise variation, as well as the top-1 relative change (T1RC) and tail relative change (TailRC) to separate the behavior of the largest singular value from that of the low-magnitude tail. TailRC is computed over the last 10% singular values.

Finally, we report the effective rank (ER) before and after finetuning, which measures the spread of the normalized singular-value distribution. A positive ΔER indicates a more spread-out spectrum after finetuning.

The results show that most models exhibit very small spectral changes, the RSD values are generally below 0.03, and the changes in effective rank are also minor. These results indicate that the singular-value distribution is nearly preserved after finetuning.

In contrast, a few models display much larger spectral changes. ConvNeXt-Base and the ConvNeXt-Large variant have substantially higher RSD, MRC, and MaxRC values, indicating stronger anisotropic rescaling of the spectrum. Notably, these models with significant spectral changes are all vision-language multimodal models.

Taken together, the singular-value analysis supports the conclusion that finetuning is spectrally constrained. For most models, finetuning leaves the pretrained singular-value spectrum almost unchanged.

Table 4: **Spectral change statistics for last-block MLP weights across ConvNeXt, ViT, and Swin models.** Most models show small relative spectral distance after finetuning, while CLIP-pretrained ConvNeXt models exhibit substantially stronger singular-value rescaling. ViT and Swin models show minimal spectral change.

Arch	Model	Layer	n_{sv}	Singular Value Change Metrics					Effective Rank		
				RSD	MRC	MaxRC	T1RC	TailRC	ER_{pre}	ER_{ft}	ΔER
ConvNeXt	CNX-Tiny (in12k)	fc1	768	0.0089	0.0089	0.2108	−0.0033	+0.0158	725.7	726.1	+0.3
		fc2	768	0.0077	0.0147	0.6706	−0.0131	+0.0150	743.9	745.2	+1.3
	CNX-Tiny (in22k)	fc1	768	0.0119	0.0137	0.1526	−0.0215	−0.0224	735.6	733.4	−2.2
		fc2	768	0.0314	0.0324	0.8587	+0.0226	+0.0151	737.3	735.6	−1.7
	CNX-Small (in22k)	fc1	768	0.0107	0.0078	0.0797	+0.0043	+0.0005	726.9	727.5	+0.7
		fc2	768	0.0188	0.0226	0.8908	+0.1028	−0.0028	739.1	737.7	−1.4
	CNX-Small (in12k)	fc1	768	0.0186	0.0201	0.3012	−0.0307	+0.0161	730.9	731.2	+0.3
		fc2	768	0.0125	0.0208	0.6847	−0.0138	+0.0285	736.6	738.8	+2.3
	CNX-Base (CLIP)	fc1	1024	0.7734	0.6639	1.0480	+0.8108	+0.4978	954.5	922.8	−31.6
		fc2	1024	1.0209	2.4158	6.9866	+0.8395	+6.3601	779.2	925.4	+146.2
	CNX-Large-MLP (CLIP)	fc1	1536	0.3304	0.2709	0.6640	+0.6640	+0.2310	1445.6	1417.6	−28.0
		fc2	1536	0.3789	0.4771	1.2856	+0.9725	+0.8735	1183.6	1208.4	+24.8
	CNX-Large (CLIP)	fc1	1536	0.0110	0.0055	0.0467	−0.0467	+0.0093	1417.3	1420.6	+3.3
		fc2	1536	0.0100	0.0057	0.0633	+0.0092	−0.0012	1471.4	1469.2	−2.2
	CNX-XLarge (in22k)	fc1	2048	0.0071	0.0042	0.0376	−0.0376	+0.0107	1855.3	1858.4	+3.0
		fc2	2048	0.0156	0.0349	0.3397	+0.0234	+0.2184	1776.2	1787.0	+10.8
	CNX-XXL (CLIP)	fc1	3072	0.0505	0.0280	0.2824	+0.2824	+0.0346	2821.1	2812.2	−9.0
		fc2	3072	0.0613	0.2843	0.5815	−0.0051	+0.3596	1673.7	1858.2	+184.6
ViT	ViT-B/16	fc1	768	0.0018	0.0013	0.0220	+0.0123	−0.0001	728.8	728.7	−0.1
		fc2	768	0.0008	0.0003	0.0035	−0.0035	−0.0002	736.2	736.2	−0.0
	ViT-L/16	fc1	1024	0.0013	0.0005	0.0117	+0.0117	+0.0004	954.9	954.9	+0.0
		fc2	1024	0.0007	0.0001	0.0019	−0.0019	+0.0002	966.4	966.4	+0.0
Swin	Swin-B	fc1	1024	0.0083	0.0086	0.0156	+0.0035	+0.0091	945.3	945.5	+0.2
		fc2	1024	0.0074	0.0080	0.0364	+0.0084	+0.0096	969.3	969.5	+0.2
	Swin-L	fc1	1536	0.0047	0.0043	0.0089	+0.0055	+0.0041	1402.6	1402.4	−0.2
		fc2	1536	0.0058	0.0051	0.0198	+0.0122	+0.0061	1447.7	1447.8	+0.0

C Spectral Gap Analysis

In the main analysis, we characterize spectral stability at both the singular-vector and subspace levels. This distinction is important because individual singular vectors can become sensitive to perturbations when neighboring singular values are closely spaced. To examine whether such closely spaced singular values occur in representative pretrained models, we analyze the spectral gaps of the fully connected (FC) weight matrices across all layers of one vision model, ConvNeXt-Base pretrained with CLIP on LAION-2B, and one language model, BERT-base-uncased.

For a singular-value sequence $\sigma_1 \geq \sigma_2 \geq \cdots$, we define the relative spectral gap at rank k as

$$g_k = \frac{\sigma_k - \sigma_{k+1}}{\sigma_k}. \tag{5}$$

Normalizing the difference by σ_k makes the measure insensitive to the overall scale of the singular values and directly quantifies their relative separation. A smaller g_k indicates that neighboring singular values are more closely spaced, making the corresponding individual singular vectors more sensitive to perturbations and rotations within the associated subspace.

Figure 7 shows the relative spectral gaps across spectral ranks for the FC weight matrices across all layers of the two representative pretrained models. In both models, the spectral gap becomes small at relatively low ranks and remains small beyond approximately rank 20. Thus, although individual leading singular vectors can provide a useful characterization of spectral alignment, their stability becomes increasingly difficult to interpret independently when neighboring singular values are closely spaced. This motivates our complementary subspace-level analysis using principal-angle similarity, which characterizes the stability of the span of multiple singular vectors without relying on the identifiability of any individual vector.

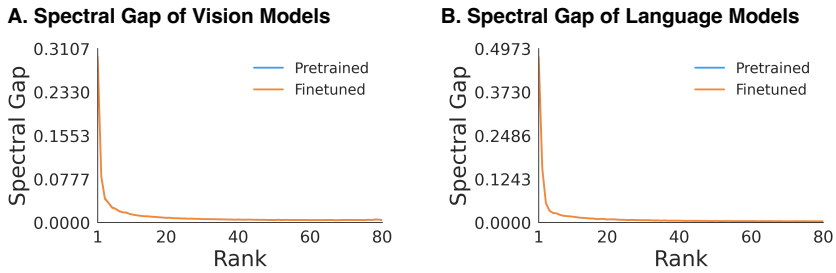


Figure 7: **Relative spectral gaps of pretrained and finetuned FC weight matrices. (A)** Relative spectral gaps across spectral ranks for the FC weight matrices across all layers of ConvNeXt-Base pretrained with CLIP on LAION-2B. **(B)** Relative spectral gaps across spectral ranks for the FC weight matrices across all layers of BERT-base-uncased.

D Spectral Stability over Broader Rank Ranges

The main text focuses on the leading spectral directions and shows that finetuning largely preserves the pretrained spectral structure, particularly in the top-ranked components. Here, we extend the analysis to a broader range of singular-value ranks and average the results across FC1 and FC2 layers, providing a unified comparison across vision and language models.

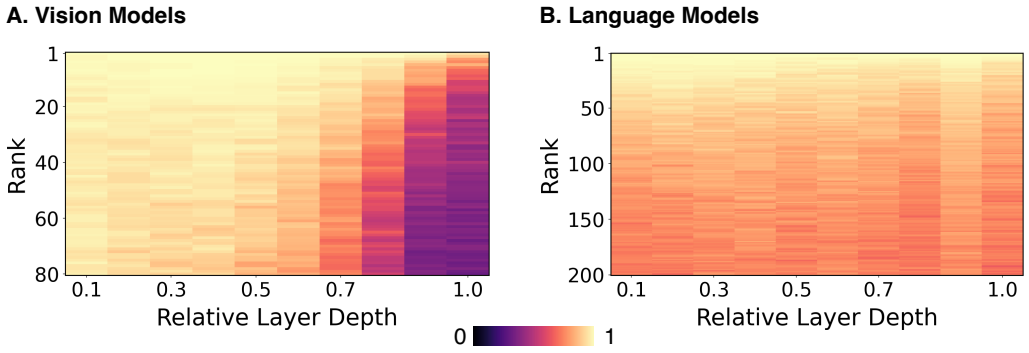


Figure 8: **Singular-vector alignment over broader spectral ranks.** Rank-dependent cosine similarity between corresponding singular vectors before and after finetuning is shown for vision and language models. Results are averaged across FC1 and FC2 layers. Leading singular vectors remain more stable after finetuning, while similarity generally decreases toward lower-ranked directions.

Figure 8 shows a consistent rank-dependent pattern across both model families. Leading singular vectors remain highly similar before and after finetuning, while similarity generally decreases toward lower-ranked directions. These results extend the main-text analysis and show that the rank-dependent organization of spectral changes persists beyond the most dominant components.

We further repeat the same analysis at the subspace level by measuring the principal-angle similarity between subspaces spanned by the leading singular vectors. As shown in Figure 9, the subspace-level results exhibit the same overall rank-dependent pattern: leading spectral subspaces remain highly stable, whereas subspace similarity generally decreases as the subspace rank increases. The consistency between vector- and subspace-level analyses further supports the robustness of the observed spectral stability over a broader range of ranks.

E Failure Cases of Spectral Stability

The main text shows that leading singular vectors of pretrained weights remain highly stable during downstream finetuning across diverse architectures and tasks. To examine whether this stability

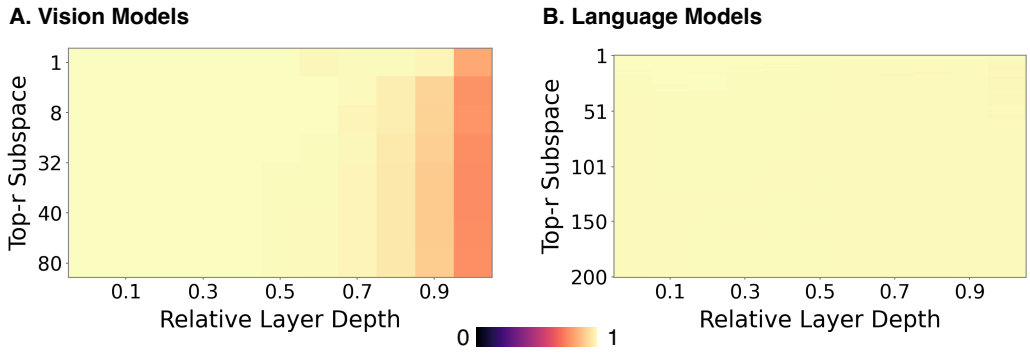


Figure 9: **Subspace similarity over broader spectral ranks.** Rank-dependent principal-angle similarity between the leading spectral subspaces before and after finetuning is shown for vision and language models. Results are averaged across FC1 and FC2 layers. Consistent with the singular-vector analysis, leading spectral subspaces remain highly stable, while subspace similarity generally decreases as the subspace rank increases.

persists when the relationship between pretraining and the downstream objective is disrupted, we construct a setting in which task structure is intentionally removed.

We isolate the role of task structure by keeping the input data fixed while randomizing the labels. Using models pretrained as described in Section 4.3, we finetune them on CIFAR10 with either true labels or randomly shuffled labels. This preserves the data distribution and training procedure while removing the semantic structure of the downstream labels.

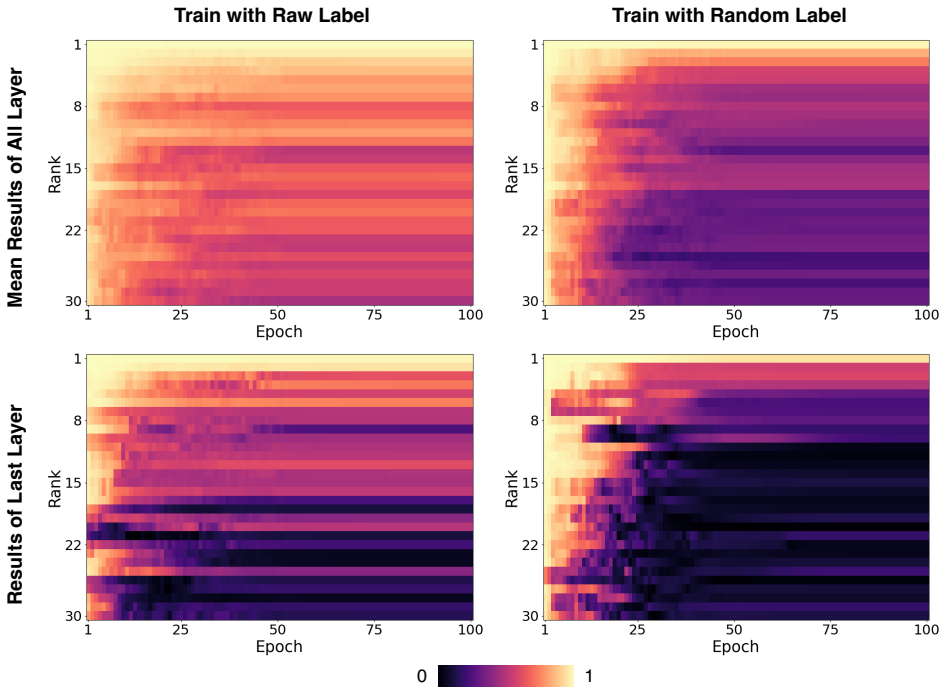


Figure 10: **Random labels reduce singular-vector stability.** Starting from the same pretrained checkpoint and using the same input data, singular vectors remain relatively stable under true labels but lose alignment more rapidly under randomly shuffled labels.

Figure 10 shows that random labels lead to substantially faster changes in individual singular vectors during training. Cosine similarity decreases more rapidly than under true labels, resulting in lower

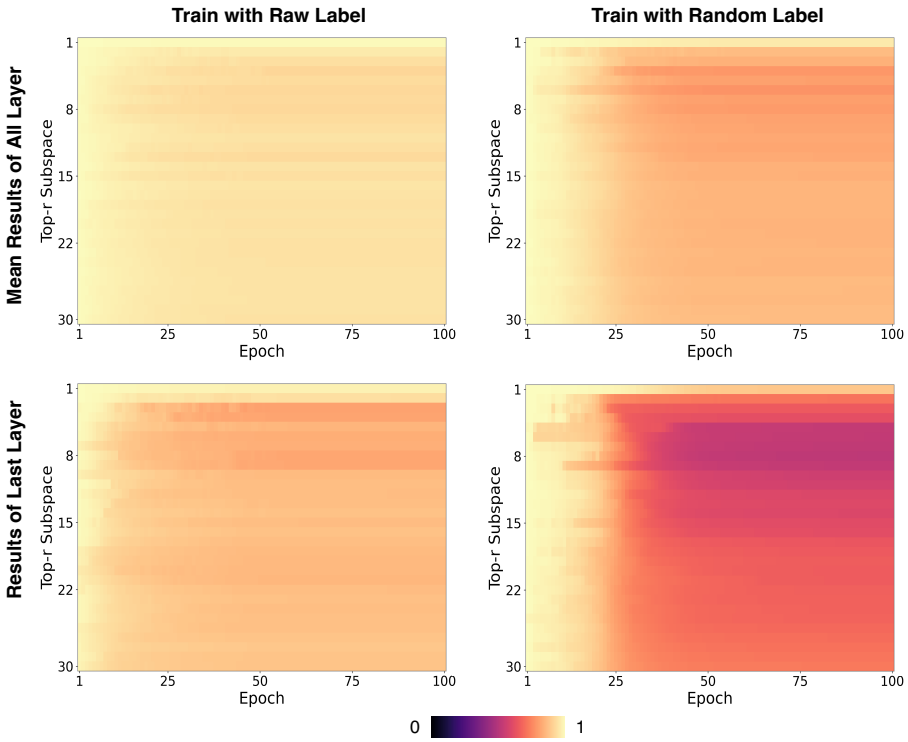


Figure 11: **Random labels reduce spectral subspace stability.** Starting from the same pretrained checkpoint and using the same input data, leading spectral subspaces remain relatively stable under true labels but lose alignment more rapidly under randomly shuffled labels, as measured by principal-angle similarity.

final similarity. We further examine the same phenomenon at the subspace level using principal-angle similarity. As shown in Figure 11, leading spectral subspaces exhibit a consistent reduction in stability under random labels, supporting the vector-level observation without relying on the stability of individual singular vectors.

Together, these results show that the spectral stability observed during standard finetuning is not simply a consequence of parameter initialization, the input data distribution, or continued optimization. Disrupting the downstream task structure through label randomization substantially reduces both singular-vector and spectral-subspace stability, providing evidence that the observed stability is related to the structure shared between pretraining and downstream learning.

F Singular-Vector Analysis in Attention Modules

The main text focuses on feed-forward (FFN) layers. To examine whether the observed spectral stability extends to attention modules, we perform the same singular-vector alignment analysis on the query, key, value, and output projection matrices.

Figures 12 and 13 show that attention projections exhibit a spectral pattern consistent with that observed in FFN layers. Across query, key, value, and output projections, leading singular vectors remain strongly aligned before and after finetuning, while alignment generally decreases toward lower-ranked components. The stability is particularly pronounced in vision models, where attention projections maintain high alignment across a broad range of leading spectral ranks.

These results extend the singular-vector analysis beyond FFN layers, suggesting that the rank-dependent spectral stability observed in the main text is not specific to a particular parameter type.

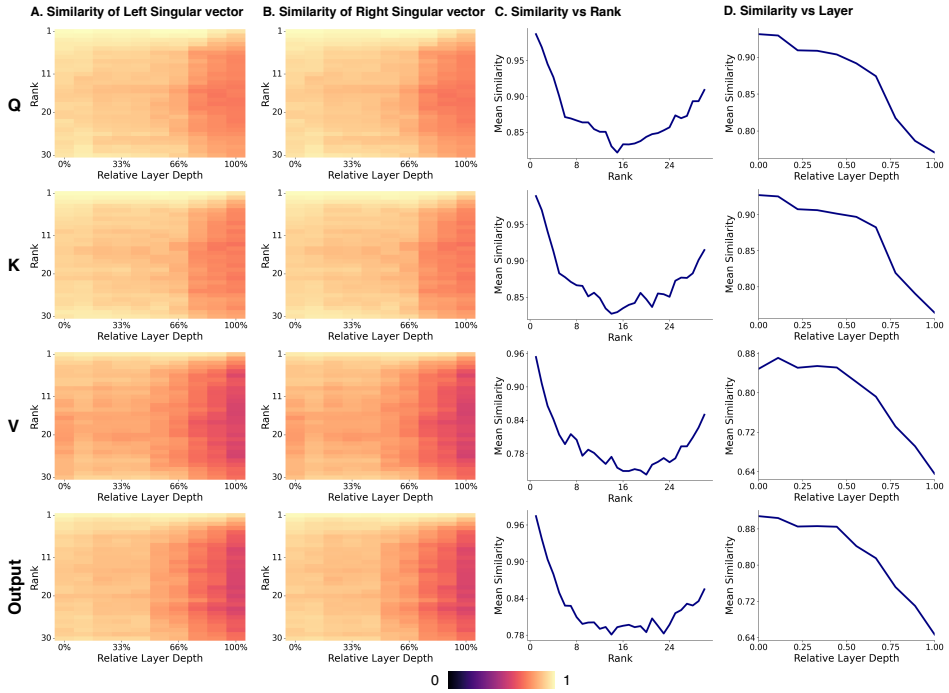


Figure 12: **Attention-module spectral alignment in vision models.** Cosine similarity between corresponding singular vectors before and after finetuning for query, key, value, and output projections.

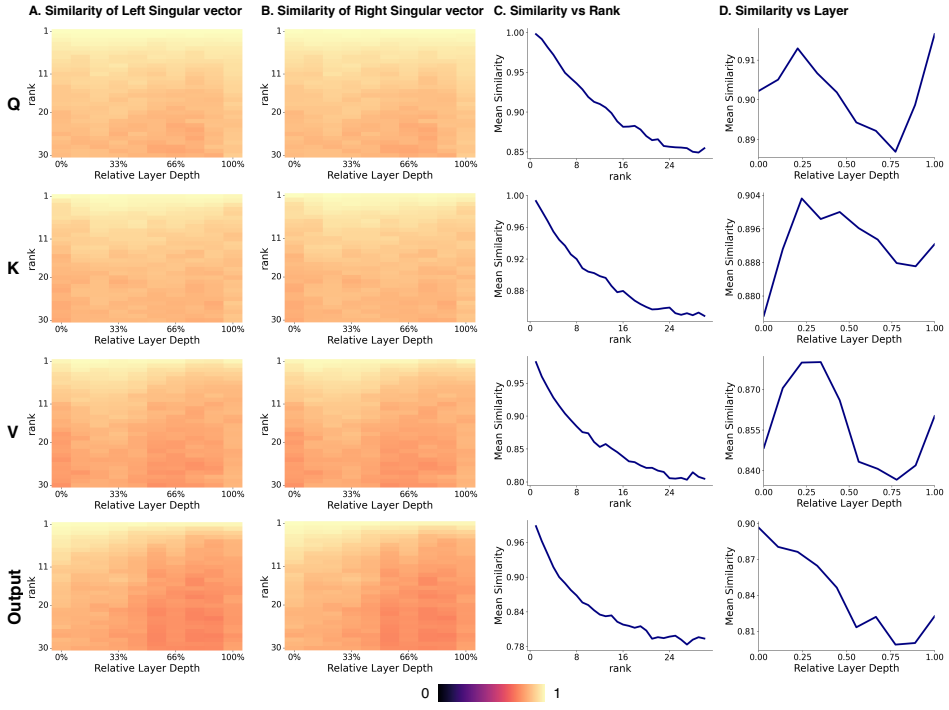


Figure 13: **Attention-module spectral alignment in language models.** Cosine similarity between corresponding singular vectors before and after finetuning for query, key, value, and output projections.

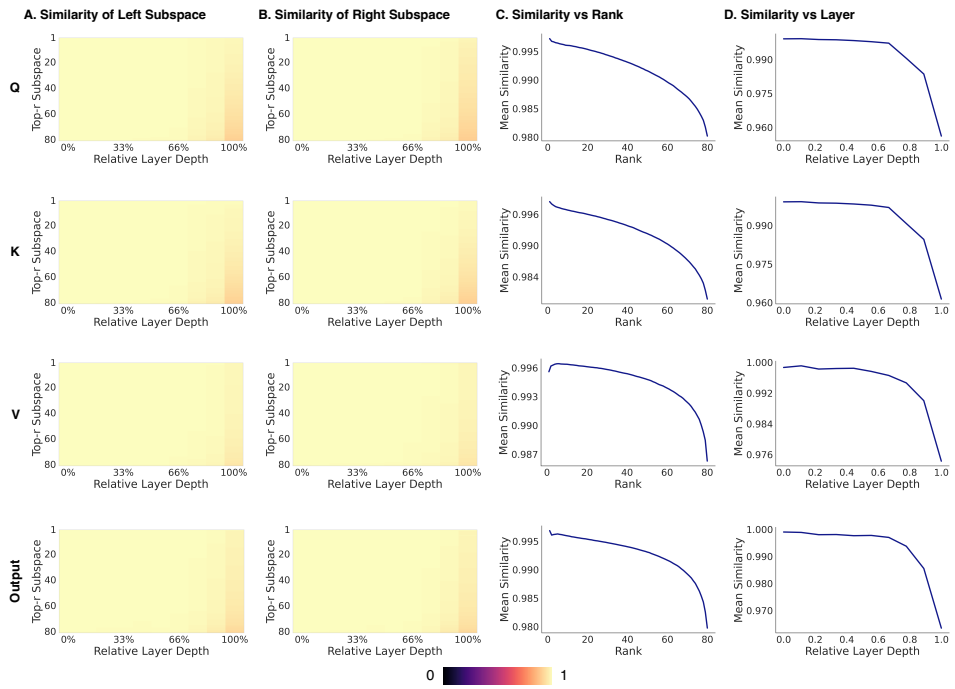


Figure 14: **Attention-module subspace stability in vision models.** Principal-angle similarity between the leading spectral subspaces of query, key, value, and output projections before and after finetuning. Leading subspaces remain highly aligned, with similarity generally decreasing as the subspace rank increases.

G Subspace Analysis in Attention Modules

Individual singular vectors can become sensitive when neighboring singular values are closely spaced. We therefore complement the vector-level analysis with a basis-invariant analysis of the corresponding spectral subspaces. For the query, key, value, and output projection matrices, we measure the principal-angle similarity between the leading k-dimensional subspaces before and after finetuning.

Figures 14 and 15 show a consistent pattern at the subspace level. The leading spectral subspaces of attention projections remain highly aligned after finetuning, while subspace similarity generally decreases as the subspace rank increases. Thus, the rank-dependent stability observed at the level of individual singular vectors is also preserved when considering the corresponding subspaces.

Together with the FFN results in the main text, these vector- and subspace-level analyses show that spectral stability extends across major parameter components of both vision and language models. In particular, attention projections preserve their dominant spectral structure during finetuning, consistent with the rank-dependent stability observed in FFN layers.

H Additional Results on Singular-Vector Swapping

The main text presents the complete singular-vector swapping analysis for FFN layers across four downstream tasks. Here, we extend the same analysis to attention modules to examine whether the observed robustness of spectral-vector swapping also holds beyond feed-forward layers.

We consider the same two swapping settings as in the main text: (i) replacing the singular vectors of a finetuned model with those from its pretrained counterpart, and (ii) swapping singular vectors between models finetuned on different downstream tasks. In both settings, the singular values are kept fixed, and we evaluate the resulting downstream performance.

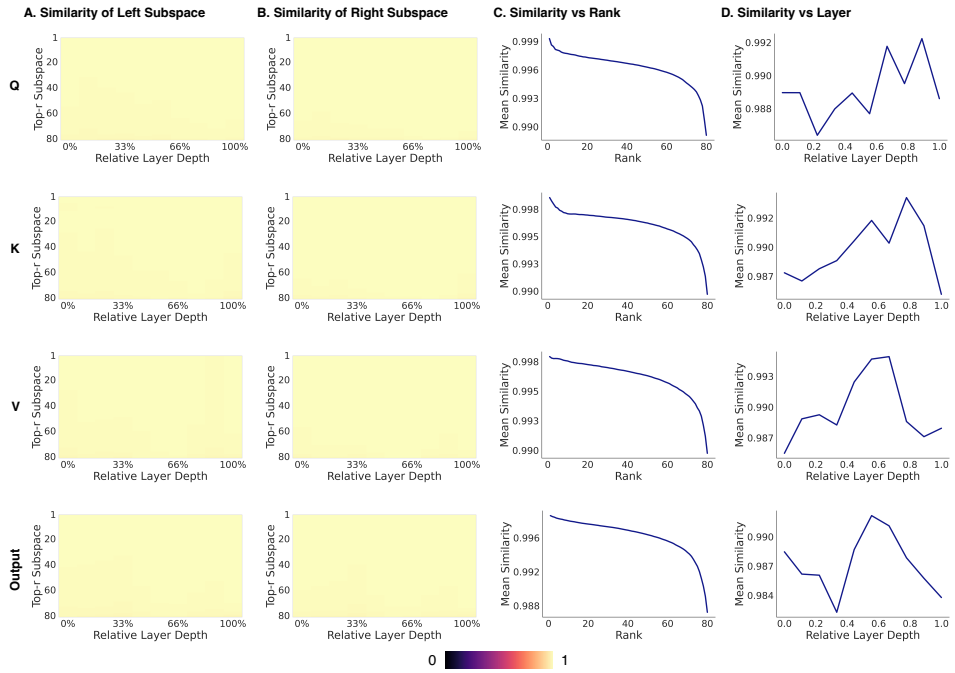


Figure 15: **Attention-module subspace stability in language models.** Principal-angle similarity between the leading spectral subspaces of query, key, value, and output projections before and after finetuning. Leading subspaces remain highly aligned, with similarity generally decreasing as the subspace rank increases.

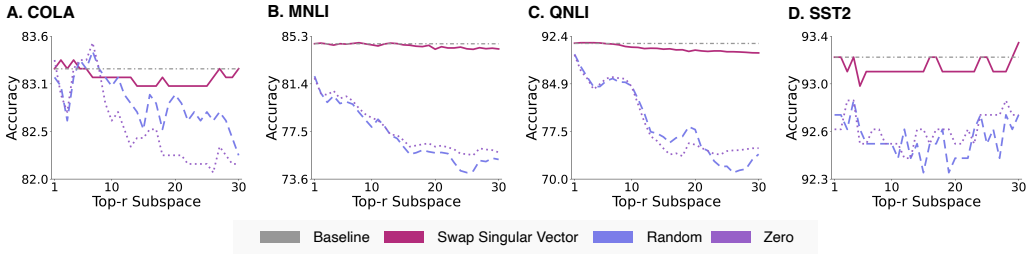


Figure 16: **Pretrained-to-finetuned singular-vector replacement in attention modules.** Replacing the singular vectors of finetuned attention projections with those from the corresponding pretrained model causes only limited performance changes, particularly for leading spectral components, consistent with the FFN results in the main text.

Figures 16 and 17 show the results for attention projections across all four tasks. The same qualitative pattern observed for FFN layers is also present in attention modules. Replacing finetuned singular vectors with their pretrained counterparts causes only limited performance changes, particularly for leading spectral components. Similarly, swapping singular vectors between models finetuned on different tasks has relatively small effects on downstream performance, although the magnitude of the change varies across tasks and spectral ranks.

These results extend the swapping analysis in the main text and show that the robustness of singular-vector replacement is not specific to FFN layers. Together with the FFN results, they provide consistent evidence that the leading spectral structure of both feed-forward and attention modules is largely preserved during finetuning and remains compatible across different downstream tasks.

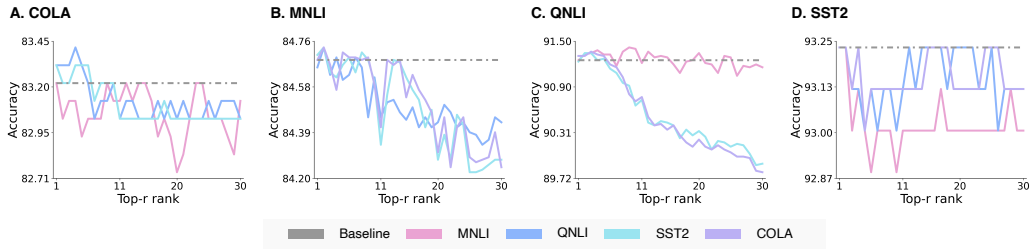


Figure 17: **Cross-task singular-vector swapping in attention modules.** Swapping singular vectors between attention projections from models finetuned on different downstream tasks causes relatively small performance changes, consistent with the cross-task swapping results for FFN layers.

I Information Content Captured by Top-K Singular Vectors

In the main text, we primarily analyzed the top-30 singular vectors of parameter matrices, and in Appendix D we extended this analysis to higher ranks. Here we provide a complementary analysis of how much variance these leading spectral components explain across different models and ranks, to contextualize the scope and implications of our findings.

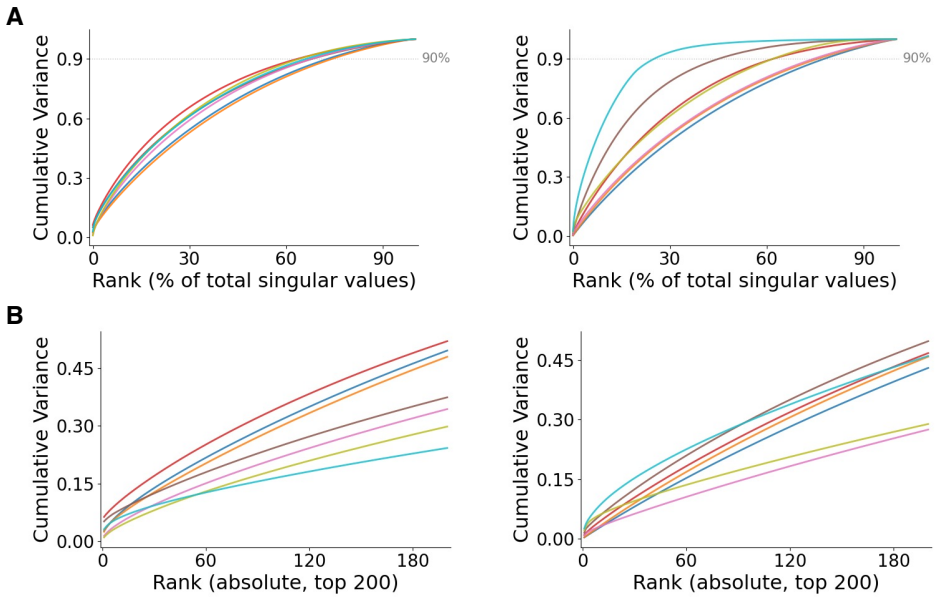


Figure 18: **Cumulative explained variance of ConvNeXt last-block FC weights.** The first and second columns correspond to **fc1** and **fc2**, respectively. **(A)** Explained variance versus relative rank. About 60% of singular values are needed to capture 90% of the Frobenius energy, with fc1 spectra slightly more concentrated than fc2. **(B)** Explained variance of the top singular values in absolute rank. The top-30 components explain about 10–15% of the variance, while the top-200 explain 30–40%. PEFT and replacement results indicate that task-relevant information is concentrated in the leading spectral components.

Figure 18 shows the cumulative explained variance as a function of rank for the last-block MLP weights (fc1 and fc2) across ConvNeXt models of varying scales. Panel (A) presents results as a function of relative rank (percentage of total singular values), while Panel (B) shows the absolute rank up to the first 200 components. Several patterns emerge. First, when measured in relative terms, approximately 60% of the total rank is required to capture 90% of the energy, with fc1 layers (first column) exhibiting slightly more concentrated spectra than fc2 (second column). Second, in absolute

terms, the top-30 singular vectors account for roughly 10–15% of the total variance, while the top-200 components explain approximately 30–40%, depending on the model.

At first glance, the relatively low explained variance of the top-30 or even top-200 singular vectors might suggest that our analysis captures only a limited view of the full parameter space. However, two key pieces of evidence demonstrate that the leading spectral directions disproportionately determine downstream task performance, even though they represent a minority of the total spectral energy. First, as shown in Table 1 and Section D, finetuning only the top-30 singular components, which explain less than 50% of total variance in most layers, achieves performance comparable to full finetuning on GLUE benchmarks. This result indicates that the bulk of the spectral mass in the tail contributes little to task-specific adaptation. The low-variance directions, despite accounting for the majority of the parameter count, appear to play a negligible role in determining model behavior on downstream tasks.

Second, our singular vector masking experiments in Section J provide direct evidence that the lower half of the spectrum is functionally inert. When we progressively mask singular vectors starting from the tail (lowest-ranked components), model performance remains virtually unchanged until a substantial portion is removed. In contrast, masking from the head (highest-ranked components) causes immediate performance degradation. This striking asymmetry reveals that although the tail singular vectors encode substantial information about the parameter matrix, this information contributes little to model generalization. The leading directions concentrate the task-relevant structure, while the tail of the spectrum serves primarily as a reservoir of unused representational capacity.

Taken together, these observations indicate that while pretrained weight matrices possess a long tail of low-variance singular values, this tail plays a limited role in transfer learning. The leading singular vectors, despite accounting for a modest fraction of total variance, capture the geometry that matters for downstream adaptation and generalization. Our focus on the top-30 to top-200 components is therefore not a limitation of our analysis, but rather a deliberate emphasis on the spectral subspace where both pretraining and finetuning concentrate their representational capacity.

J Sensitivity of Masking Order

To investigate the role of spectral rank ordering, we compare two complementary masking strategies: (i) masking from the top rank downward, and (ii) masking from the bottom rank upward.

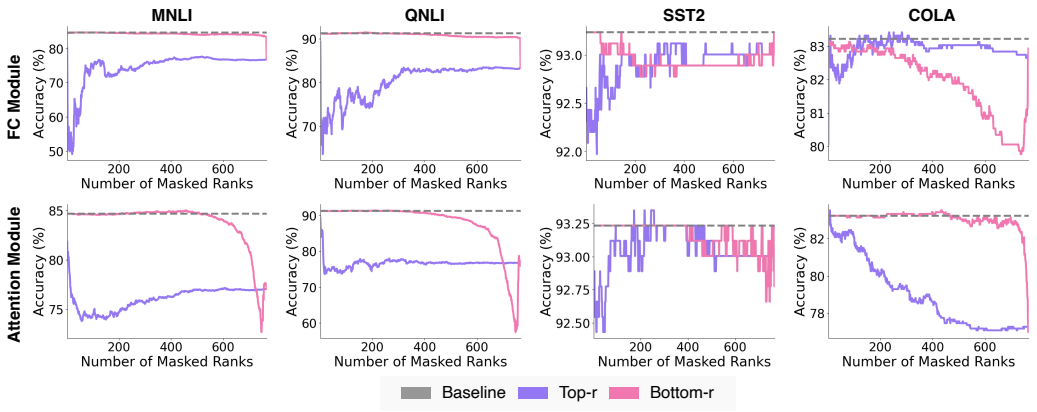


Figure 19: **Masking order comparison across tasks and module types.** Performance under top-down masking (from leading ranks) versus bottom-up masking (from trailing ranks) on four datasets. Results are shown for both feed-forward (FC) and attention modules.

Figure 19 shows that these two strategies yield markedly different behaviors. When masking begins from the top ranks, performance degrades rapidly, indicating that leading singular components carry critical task-relevant information. In contrast, masking from the bottom ranks preserves performance

remarkably well, in some tasks, over 90% of the trailing singular vectors can be masked with negligible impact.

This asymmetry reveals a key insight: although the top-K singular vectors account for a relatively small fraction of the total spectral energy (as shown in Appendix I), they disproportionately determine downstream task performance. While these leading components may explain less than 50% of the parameter variance, they capture the directions most relevant for generalization. Conversely, the majority of the spectral basis, despite representing most of the parameter mass, remains largely task-agnostic and can be safely discarded without affecting performance. This reinforces our central finding that task-specific information is concentrated in a small number of leading spectral components, and that the pretrained spectral basis provides a transferable coordinate system where only a few directions require adaptation.

K Additional Results on the Effect of Pretraining Data Scale on Spectral Stability

Pretraining setup. All pretraining experiments are conducted on CIFAR-10. We train Masked Autoencoder (MAE) models from scratch using varying amounts of training data, while fixing the validation set to the last 5,000 samples of the CIFAR-10 test set. The best checkpoint is selected based on validation loss. Images are resized to 32×32 and divided into 4×4 patches, yielding 64 patches per image. We use a masking ratio of 60%. The encoder follows a ViT-Tiny configuration with embedding dimension 192, 6 Transformer blocks, 3 attention heads, and 2D sinusoidal positional encoding. The decoder uses an embedding dimension of 64, 2 Transformer blocks, and 4 attention heads, followed by a linear projection to predict pixel values for each patch with dimension $4 \times 4 \times 3 = 48$. The reconstruction objective is per-patch MSE loss. We train for 200 epochs with batch size 64 using AdamW with learning rate 1.5×10^{-3} and weight decay 10^{-4}, together with cosine annealing, gradient clipping at 1.0, and mixed-precision training using AMP and GradScaler with TF32 acceleration. Validation is performed every epoch, and model weights are saved whenever the validation loss reaches a new minimum.

Post-training scenarios. We conduct two types of post-training experiments starting from the best pretrained checkpoint at each data scale. For same-task continued training, we further train the MAE model on the first 1,000 samples of the CIFAR-10 test set, using the same architecture and reconstruction objective as in pretraining. We train for 10 epochs with batch size 64 using AdamW with learning rate 3×10^{-3} and weight decay 10^{-4}, together with cosine annealing and gradient clipping at 1.0. For classification finetuning, we reuse the pretrained encoder and add a classification head consisting of global average pooling over the encoder patch tokens followed by a linear layer. Only the encoder parameters are loaded from the pretrained checkpoint, while the classification head is randomly initialized. We finetune on the first 3,000 samples of the CIFAR-10 test set for 10 epochs using cross-entropy loss and AdamW with learning rate 1.5×10^{-3}, weight decay 10^{-4}, cosine annealing, and gradient clipping at 1.0.

In the main text, we illustrate the post-training evolution of singular-vector similarity for two representative pretraining data scales, 1,000 and 50,000 samples. Here, we provide the complete results across a broader range of pretraining data scales.

Figure 20 shows the singular-vector similarity between the current model and its pretrained initialization throughout post-training. Across all data scales and both post-training settings, similarity generally decreases as adaptation proceeds, indicating that post-training progressively changes the pretrained spectral directions. However, the magnitude of these changes depends on the amount of pretraining data. Models pretrained on larger datasets consistently maintain higher singular-vector similarity throughout post-training and exhibit higher similarity at the end of adaptation.

This pattern is observed not only for the two representative data scales shown in the main text, but also across intermediate pretraining scales. While the magnitude of the effect varies across settings, the overall trend remains consistent, providing additional evidence that larger-scale pretraining is associated with greater stability of individual spectral directions under subsequent task and data shifts.

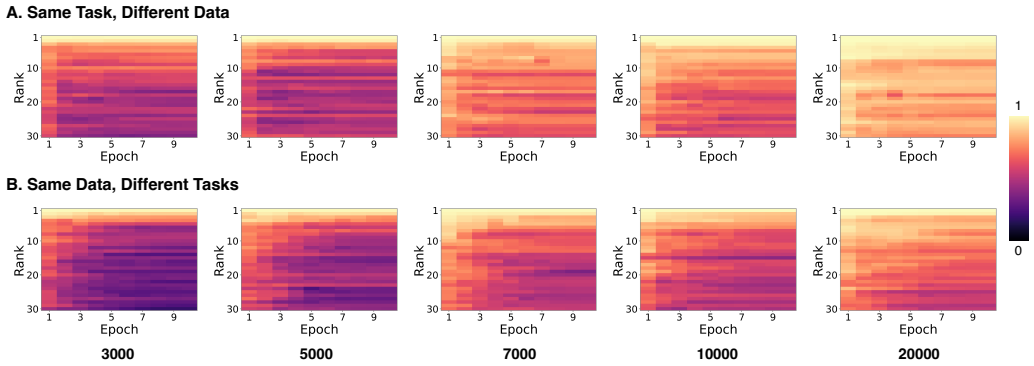


Figure 20: **Additional results across pretraining data scales at the singular-vector level. (A)** Post-training on the same task with a different data subset. **(B)** Post-training with a different task on the same data source. In both settings, models pretrained on larger datasets maintain higher singular-vector similarity to their pretrained initialization throughout post-training.

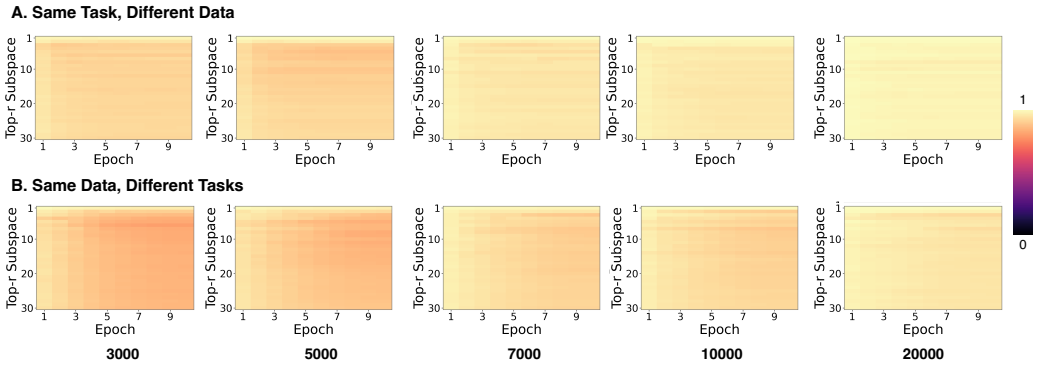


Figure 21: **Additional results across pretraining data scales at the subspace level. (A)** Principal-angle similarity of the leading spectral subspaces during post-training on the same task with a different data subset. **(B)** Principal-angle similarity during post-training with a different task on the same data source. In both settings, models pretrained on larger datasets maintain higher similarity between their leading spectral subspaces and the corresponding pretrained subspaces.

We further examine the same effect at the subspace level using principal-angle similarity. Figure 21 shows the similarity between the leading spectral subspaces of the current model and its pretrained initialization throughout post-training. The results are consistent with the singular-vector analysis: larger pretraining datasets lead to higher subspace similarity throughout adaptation in both post-training settings. Thus, the effect of pretraining data scale is not limited to the alignment of individual singular vectors, but also extends to the stability of the corresponding leading spectral subspaces.

Together, these results provide a more comprehensive view of how pretraining data scale relates to spectral stability. Increasing the amount of pretraining data is associated with greater stability at both the singular-vector and subspace levels under subsequent changes in task or data distribution.

L Implementation Details for Spectral Parameter-Efficient Transfer

This section provides implementation details for the spectrally restricted parameter-efficient finetuning experiments in Section 4.4. The goal of these experiments is to test whether downstream adaptation can be performed effectively by updating only a small number of coordinates in the singular-vector basis of the pretrained model.

Base model and task setup. All experiments use `microsoft/deberta-v3-base` as the pretrained backbone. For the controlled ablation experiments in Figure 5, we use MNLI as the representative downstream task. MNLI is a three-way natural language inference task with labels `entailment`, `contradiction`, and `neutral`. The maximum sequence length is set to 256. Unless otherwise specified, the random seed is fixed to 42.

Spectral parameterization. We apply the spectrally restricted adapter to the feed-forward module weights of DeBERTa. For each selected pretrained weight matrix, we compute its singular value decomposition $W_0 = U\Sigma V^\top$. The pretrained singular vectors U and V are kept fixed throughout finetuning, and only a small block of spectral coefficients is made trainable. Specifically, we select a contiguous spectral interval $[k, m)$, where k is the starting singular-coordinate index and m is the ending index. The effective trainable rank is therefore $r = m - k$. For example, $k = 0$ and $m = 300$ means that the method adapts the first 300 pretrained spectral coordinates. This restricts finetuning to a low-dimensional subspace defined by the pretrained spectral basis.

For DeBERTa-v3-base, the relevant FFN dimension is $d = 768$, so the valid spectral index range is bounded by $0 \le k < m \le 768$. In the rank-sweep experiments, we vary m in increments of 50.

Target layers and trainable parameters. By default, the spectral adapter is applied only to the last two DeBERTa encoder blocks, i.e., blocks 10 and 11. Within each selected block, we modify the FFN matrices `intermediate.dense` and `output.dense`. All other pretrained backbone parameters are frozen.

Ablation experiments in Figure 5. Figure 5 contains three controlled experiments on MNLI. These experiments are designed to evaluate the effects of spectral rank, layer depth, and the location of the selected spectral block.

- **Experiment 1: rank-budget sweep.** We fix $k = 0$ and vary $m \in \{50, 100, 150, \ldots, 700\}$. Since $k = 0$, the effective trainable rank is $r = m$. This experiment tests whether increasing the number of rank improves downstream performance.
- **Experiment 2: layer-location comparison.** We fix the spectral range to $k = 0$ and $m = 200$, and vary the layers to which the adapter is applied. This experiment tests whether adapting deeper layers leads to better generalization performance.

For these MNLI ablations, we train for 3 epochs with batch size 16, learning rate 1×10^{-4}, weight decay 0.0, and warmup ratio 0.06. We use the full training set. The default target blocks are $[10, 11]$ unless the layer-location experiment specifies otherwise.

GLUE benchmark setting. For the full GLUE comparison in Table 1, we use the same DeBERTa-v3-base backbone and apply the spectral adapter to the final two FFN blocks. For all GLUE tasks, we adapt the leading 300 pretrained spectral coordinates in the FFN matrices of layers 10 and 11. The batch size is fixed to 16, weight decay is set to 0.0, the warmup ratio is 0.06, and the random seed is 42 for all tasks. The maximum sequence length is 256.

Table 5 reports the task-specific hyperparameters, namely the number of training epochs and the learning rate.

Table 5: Task-specific hyperparameters for the GLUE experiments in Table 1.

Task	Epochs	Learning Rate
SST-2	20	1×10^{-4}
MRPC	12	2×10^{-4}
MNLI	5	2×10^{-4}
CoLA	30	3×10^{-4}
QNLI	10	2×10^{-4}
QQP	10	2×10^{-4}
RTE	20	2×10^{-4}
STS-B	10	2×10^{-4}

PEFT baselines. We compare our spectral-subspace-based adaptation with several representative parameter-efficient fine-tuning (PEFT) methods, including PiSSA (Meng et al., 2024), LoRA-K, LoRA-G (Hu et al., 2022), AdaLoRA (Zhang et al., 2023a), DoRA (Liu et al., 2024), HAdapter (Jeong et al., 2026), PAdapter (Newman et al., 2021), SVFT (Lingam et al., 2024) and BitFit (Zaken et al., 2022). The trainable parameter budgets range from 0.003M to 1.33M, compared with 184M parameters for full fine-tuning.

Optimization and implementation details. The SVD of each selected pretrained weight matrix is computed once before downstream training. The singular vectors are then frozen, and finetuning optimizes only the selected spectral coefficients and the task-specific prediction head. We use the AdamW optimizer with a linear learning-rate schedule and warmup, where the warmup phase occupies 6% of the total training steps. No weight decay is used.

For robustness when downloading pretrained checkpoints or datasets, we allow up to 5 retries. The base waiting time between retries is 8 seconds. These retry parameters only affect data and checkpoint loading and do not change the training procedure.

Computational resources. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24GB of memory. The CPU is an AMD EPYC 7T83 64-Core Processor.

Summary. These implementation choices make the PEFT experiment a direct test of the paper's central hypothesis. The pretrained singular-vector basis is fixed, and downstream learning is restricted to a small number of spectral coordinates in the final FFN layers. The competitive GLUE performance obtained under this constraint indicates that a compact subset of pretrained spectral directions already provides an effective basis for parameter-efficient transfer.