REVISITING SINGULAR STRUCTURE IN DATA-FREE MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent data-free model merging methods increasingly exploit the singular structure of task updates, yet they often couple singular-vector support and singular-value weighting within a single spectral operator. We revisit this design choice and study how different components of the singular structure should actually be used for merging. Under a common covariance-weighted framework, we disentangle task support, within-support anisotropy, and task-layer scale. Our analysis shows that right singular vectors provide a useful task-relevant support, while singular values are more effective when compressed into a scalar task-layer scale than when used for direction-wise reweighting. In particular, a scale-only inverse spectral response consistently outperforms its anisotropic counterpart under matched support and total scale. Motivated by this finding, we propose SOS-Merge, which represents each task operator as a right-singular projector scaled by a fractional inverse spectral moment. We further develop a data-free calibration rule for the shared spectral exponent, achieving performance close to post-hoc grid search without validation data. Across three CLIP ViT backbones and 8-, 14-, and 20-task merging settings, SOS-Merge consistently outperforms recent data-free baselines, with its advantage becoming more pronounced as more tasks are merged.

1 INTRODUCTION

Modern foundation models are often adapted into multiple task-specific experts, but maintaining a separate model for every task is costly in storage, deployment, and inference. Model merging offers an alternative by combining experts fine-tuned from a common initialization into a single model, preferably without access to the original task data (Wortsman et al., 2022; Ilharco et al., 2023; Yang et al., 2026). Early data-free methods build on the task-vector view, where fine-tuned checkpoints are represented as parameter updates from the pretrained model and combined directly in weight space (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023). Methods such as TIES-Merging (Yadav et al., 2023) and Model Breadcrumbs (Davari & Belilovsky, 2024) improve this process by resolving conflicts or sparsifying task updates. Recent work has moved beyond coordinate-wise operations and started to exploit the matrix structure of task updates. For a matrix update $\Delta = U\Sigma V^\top$, spectral and subspace methods use singular-vector subspaces, spectral weighting, covariance proxies, or spectral smoothing to reduce interference (Marczak et al., 2025; Gargiulo et al., 2025; Xu et al., 2026). A common spectral construction combines the singular-vector subspace with weights derived from the singular values. This couples several effects. The chosen subspace determines where the task acts, while the singular values determine both the relative weighting of directions within that subspace and the overall strength of the resulting operator. Consequently, an improvement from spectral weighting does not reveal whether the benefit comes from the subspace itself, from anisotropic weighting within the subspace, or from the aggregate strength of the spectrum. These effects have largely been evaluated together rather than isolated.

We disentangle these roles through a controlled analysis. We first verify that the right singular subspace provides a meaningful proxy for task-relevant input structure. We then separate the within-subspace anisotropy induced by the singular values, namely their relative weights across directions, from the total spectral mass. Comparing operators with the same support and the same total spectral mass isolates the effect of this anisotropy on merging. Across all controlled comparisons, we find retaining the anisotropic weights yields lower merging performance than collapsing the spectrum

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.

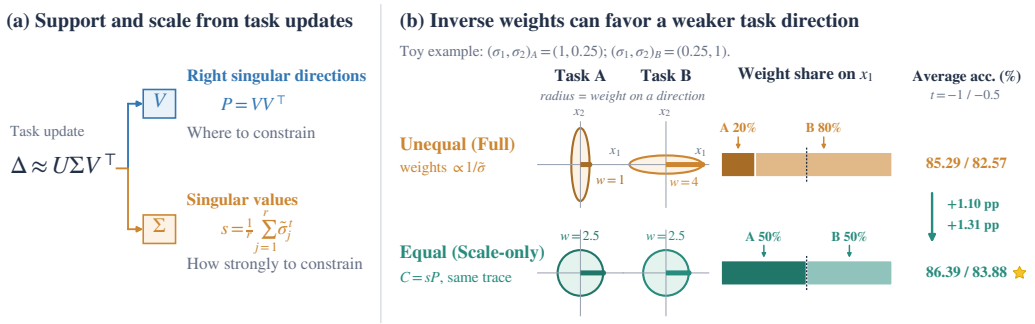


Figure 1: **Separating task support from spectral strength in model merging.** (a) The retained right-singular subspace defines the task support $P = VV^\top$, while the singular values determine its aggregate spectral strength. (b) Direction-specific inverse weighting can overemphasize a task along a shared direction; in the toy example, Task B receives 80% of the weight on x_1. Scale-only replaces the direction-specific weights by a single scalar under matched support and total spectral mass, giving equal 50% shares. Across the nine settings in Tab. 3, Scale outperforms Full by 1.10 and 1.31 points for $t = -1$ and $t = -0.5$, respectively.

into a single scalar. This suggests that the useful information in the singular values is captured more effectively by aggregate spectral strength than by direction-specific variation.

This finding leads to SOS-MERGE (Scale-Only Spectral Merging). SOS-MERGE keeps the right-singular subspace as task support and summarizes the retained spectrum with a single scalar weight. We derive this weight from a fractional inverse spectral response and calibrate its strength directly from the task checkpoints, requiring neither task examples nor validation accuracy. In this way, the spectrum controls the strength of the task operator while the geometry of its support remains isotropic. For classification tasks, we further consider a lightweight class-aware refinement. We summarize our contributions as follows:

- **Finding.** We separate the roles of task support, within-support anisotropy, and overall spectral strength in data-free model merging. Our controlled analysis shows that right singular vectors provide task-relevant support, while singular values are more effective when compressed into a scalar strength than when used for direction-wise reweighting.
- **Method.** Based on this finding, we propose SOS-MERGE, which decouples right-singular support from spectral strength and summarizes the retained singular values with a fractional inverse spectral response. Its shared exponent is calibrated from checkpoint spectra alone, without validation data.
- **Strong data-free performance.** Across three CLIP ViT backbones and 8, 14, and 20 task merging settings, SOS-MERGE achieves an average accuracy of 86.92%, outperforming the strongest data-free baseline, SWUDI-A, by 1.16 percentage points on average and ranking first in 8 of 9 settings.

2 DISENTANGLING THE SINGULAR STRUCTURE OF TASK UPDATES

2.1 A COMMON COVARIANCE-WEIGHTED MERGING FRAMEWORK

Let θ_0 denote a pretrained model and $\theta_i = \theta_0 + \Delta\theta_i$ the expert model fine-tuned for task i. Consider a matrix parameter at layer ℓ, with pretrained weight $W_{0\ell}$, task update $\Delta_{i\ell}$, and merged update M_ℓ. For an input activation x from task i for layer ℓ, matching the merged and expert layer outputs gives

$$\mathbb{E}_{x\sim\mathcal{D}_i} \|(M_\ell - \Delta_{i\ell})x\|_2^2 = \left\|(M_\ell - \Delta_{i\ell})(C_{i\ell}^{\text{act}})^{1/2}\right\|_F^2, \tag{1}$$

where $C_{i\ell}^{\text{act}} = \mathbb{E}_{x\sim\mathcal{D}_i}[xx^\top]$ is the input activation covariance. In the data-free setting, $C_{i\ell}^{\text{act}}$ is unavailable and must be replaced by an operator constructed from the model checkpoints.

This motivates the following common layer-wise objective:

$$\min_{M_\ell} \sum_{i=1}^{K} w_i \left\|(M_\ell - \Delta_{i\ell})C_{i\ell}^{1/2}\right\|_F^2 + \left\|M_\ell C_{0\ell}^{1/2}\right\|_F^2 + \lambda\|M_\ell\|_F^2, \tag{2}$$

Table 1: **Right singular subspaces align with task activation subspaces.** The task-update support $V_{i\ell}V_{i\ell}^\top$ shows substantially higher overlap and smaller principal angles than a dimension-matched random support. Values are averaged over the analyzed task–layer pairs.

Support proxy	Average overlap ↑	Average angle (°) ↓
Random support	0.06	77.59
Task-update support $V_{i\ell}V_{i\ell}^\top$	**0.35**	**56.48**

where $w_i \geq 0$ is the weight of task i, $C_{i\ell} \succeq 0$ is its task operator, $C_{0\ell} \succeq 0$ is a fixed pretrained-model operator, and $\lambda \geq 0$ is a ridge coefficient. The objective has the closed-form solution

$$M_\ell^\star = \left(\sum_i w_i \Delta_{i\ell} C_{i\ell}\right)\left(\sum_i w_i C_{i\ell} + C_{0\ell} + \lambda I\right)^{-1}. \tag{3}$$

Throughout the analysis, we use uniform task weights $w_i = 1/K$ and keep $C_{0\ell}$ and λ fixed. Thus, all comparisons use the same objective and solver and differ only in the construction of $C_{i\ell}$.

For a rank-$r_{i\ell}$ truncated SVD $\Delta_{i\ell} \approx U_{i\ell}\Sigma_{i\ell}V_{i\ell}^\top$, where $V_{i\ell}$ contains the retained right singular vectors, we consider spectral task operators of the form

$$C_{i\ell} = V_{i\ell} g(\Sigma_{i\ell}) V_{i\ell}^\top, \tag{4}$$

where $g(\Sigma_{i\ell})$ is a nonnegative diagonal spectral response. In this parameterization, $V_{i\ell}$ determines the subspace on which the task operator acts, while $g(\Sigma_{i\ell})$ assigns weights to directions within that subspace. These weights affect two quantities simultaneously: their relative variation determines the within-subspace anisotropy, while their overall magnitude determines the total spectral mass of the operator. The following analysis separates these effects to identify which parts of the singular structure are useful for merging.

2.2 Right Singular Subspaces Provide a Task-Relevant Support Proxy

We first isolate the support of the spectral operator. For the truncated task update $\Delta_{i\ell} \approx U_{i\ell}\Sigma_{i\ell}V_{i\ell}^\top$, we remove the singular-value weights and retain only the projector onto the right singular subspace,

$$P_{i\ell} = V_{i\ell}V_{i\ell}^\top. \tag{5}$$

Since $P_{i\ell}$ acts in the input space of layer ℓ, we examine whether its subspace aligns with the dominant input-activation subspace induced by the same task. This analysis uses held-out task inputs only as a diagnostic. They are not used by the merging method or for model selection. Let $\{x_n\}_{n=1}^N$ denote input activations to layer ℓ from task i, and define their empirical covariance as

$$\widehat{C}_{i\ell}^{\mathrm{act}} = \frac{1}{N}\sum_{n=1}^N x_n x_n^\top = A_{i\ell}\Lambda_{i\ell}A_{i\ell}^\top. \tag{6}$$

Let $Q_{i\ell} \in \mathbb{R}^{d_\ell^{\mathrm{in}} \times r_{i\ell}}$ contain the leading $r_{i\ell}$ eigenvectors of $\widehat{C}_{i\ell}^{\mathrm{act}}$. We quantify the alignment between the task-update and activation subspaces using

$$\mathrm{Overlap}(V_{i\ell}, Q_{i\ell}) = \frac{1}{r_{i\ell}}\left\|V_{i\ell}^\top Q_{i\ell}\right\|_F^2, \qquad \vartheta_j = \arccos\sigma_j\left(V_{i\ell}^\top Q_{i\ell}\right), \tag{7}$$

where $\{\vartheta_j\}_{j=1}^{r_{i\ell}}$ are the principal angles. Larger overlap and smaller principal angles indicate stronger subspace alignment. As a reference, we compare against dimension-matched random subspaces and average the random results over 16 draws.

Tab. 1 shows a clear alignment between the right singular subspace and the task activation subspace. Compared with the random reference, the average overlap increases from 0.06 to 0.35, while the average principal angle decreases from $77.59°$ to $56.48°$. The right singular subspace therefore provides a meaningful checkpoint-derived proxy for task-relevant input structure even before any singular-value weighting is introduced. Having isolated the support, we next examine how the singular values should be used within it.

Table 2: **Direct versus inverse spectral responses.** Macro test top-1 accuracy (%). Inverse responses win in all nine settings at both tested magnitudes.

		Strong		Moderate	
Backbone	K	$t=1$	$t=-1$	$t=0.5$	$t=-0.5$
ViT-B/32	8	54.89	**86.05**	64.05	**84.21**
	14	60.18	**81.07**	64.31	**78.31**
	20	58.36	**76.63**	61.50	**73.17**
ViT-B/16	8	61.41	**89.05**	68.78	**87.17**
	14	64.60	**85.13**	68.52	**81.70**
	20	61.84	**79.74**	64.79	**75.94**
ViT-L/14	8	70.52	**91.91**	77.35	**90.60**
	14	71.33	**89.85**	76.07	**87.42**
	20	67.41	**88.17**	71.50	**84.58**
Average	–	63.39	**85.29**	68.54	**82.57**

Table 3: **Scale-only versus anisotropic inverse spectra.** Macro test top-1 accuracy (%). Scale beats the trace-matched Full variant in every setting.

		Strong ($t=-1$)			Moderate ($t=-0.5$)		
Backbone	K	Shape	Full	Scale	Shape	Full	Scale
ViT-B/32	8	70.47	86.05	**86.16**	74.04	84.21	**84.91**
	14	67.69	81.07	**82.35**	69.42	78.31	**79.97**
	20	63.53	76.63	**79.31**	64.98	73.17	**75.73**
ViT-B/16	8	75.72	89.05	**89.27**	78.22	87.17	**87.74**
	14	71.83	85.13	**86.05**	73.44	81.70	**83.21**
	20	66.99	79.74	**82.32**	68.28	75.94	**78.02**
ViT-L/14	8	83.32	91.91	**92.30**	84.83	90.60	**90.98**
	14	79.05	89.85	**90.27**	80.89	87.42	**88.27**
	20	73.75	88.17	**89.48**	75.80	84.58	**86.10**
Average	–	72.48	85.29	**86.39**	74.43	82.57	**83.88**

2.3 Direct versus Inverse Spectral Responses

Having fixed the right-singular support, we next examine how singular values should affect the weights assigned within that support. A direct spectral response gives larger weights to directions with larger singular values, as in covariance constructions based on $\Delta_{i\ell}^\top \Delta_{i\ell} \approx V_{i\ell}\Sigma_{i\ell}^2 V_{i\ell}^\top$. We compare this choice with an inverse response that emphasizes directions associated with smaller retained singular values.

Let $\widetilde{\sigma}_{i\ell j} = \max(\sigma_{i\ell j}, \epsilon)$, $\widetilde{\Sigma}_{i\ell} = \mathrm{diag}_j\left(\widetilde{\sigma}_{i\ell j}\right)$, where ϵ is a small numerical floor for negative powers. We define $C_{i\ell}^{\mathrm{full}}(t) = V_{i\ell}\widetilde{\Sigma}_{i\ell}^{\,t} V_{i\ell}^\top$, using the same exponent t for all tasks and layers. Positive t preserves the ordering of the singular values, assigning more weight to directions with larger singular values, while negative t reverses this ordering. We compare $t = \pm 1$ and $t = \pm 0.5$, keeping the right-singular support and exponent magnitude fixed within each pair.

Tab. 2 shows a clear preference for inverse responses. For $|t| = 1$, the average accuracy increases from 63.39% with $t = 1$ to 85.29% with $t = -1$. For $|t| = 0.5$, the corresponding increase is from 68.54% to 82.57%. The inverse response performs better in all 9 settings for both exponent magnitudes. However, changing the response sign also changes the total spectral mass of the operator. This comparison therefore cannot determine whether the gain comes from the direction-specific inverse weights or from their aggregate magnitude.

2.4 Does Spectral Information Need to Be Direction-Specific?

To distinguish the two coupled effects, we hold the right-singular support and the total spectral mass fixed while varying the within-subspace anisotropy. For a given exponent t, define the average spectral response

$$\mu_{i\ell}(t) = \frac{1}{r_{i\ell}} \mathrm{tr}\left(\widetilde{\Sigma}_{i\ell}^{\,t}\right) = \frac{1}{r_{i\ell}} \sum_{j=1}^{r_{i\ell}} \widetilde{\sigma}_{i\ell j}^{\,t}. \tag{8}$$

We then construct two controlled variants of the full spectral operator:

$$C_{i\ell}^{\mathrm{shape}}(t) = \frac{1}{\mu_{i\ell}(t)} C_{i\ell}^{\mathrm{full}}(t), \qquad C_{i\ell}^{\mathrm{scale}}(t) = \mu_{i\ell}(t) P_{i\ell}. \tag{9}$$

The Shape operator $C_{i\ell}^{\mathrm{shape}}(t)$ removes the overall magnitude while preserving the relative weights across singular directions. The Scale operator $C_{i\ell}^{\mathrm{scale}}(t)$ does the opposite. It removes the direction-specific variation while preserving the average spectral response. In particular,

$$\mathrm{tr}\, C_{i\ell}^{\mathrm{full}}(t) = \mathrm{tr}\, C_{i\ell}^{\mathrm{scale}}(t) = r_{i\ell}\mu_{i\ell}(t), \tag{10}$$

whereas $C_{i\ell}^{\mathrm{shape}}(t)$ has unit average eigenvalue on the retained support. The Full–Scale comparison therefore isolates the effect of within-subspace anisotropy, since the two operators share both the same support and the same total spectral mass.

Tab. 3 shows that this anisotropy is not beneficial. Scale outperforms the trace-matched Full operator in every backbone–task-count setting for both tested inverse responses, improving the average accuracy by 1.10 points at $t = -1$ and 1.31 points at $t = -0.5$. Removing the overall spectral magnitude is substantially more damaging: Shape trails Scale by 13.91 and 9.45 points on average for the two responses. These results show that the benefit of the singular values is captured more effectively by their aggregate spectral strength than by their relative variation across directions.

3 A Scale-Only Spectral Merger

3.1 Scale-Only Task Operator

The analysis in Sec. 2 shows that the right-singular subspace provides useful task support, while the singular values are more effective when summarized by their aggregate spectral strength than when used for direction-specific weighting. We therefore define a scale-only task operator

$$C_{i\ell} = s_{i\ell} P_{i\ell}, \qquad s_{i\ell} = G_{i\ell}(q_i) = \frac{1}{r_{i\ell}} \sum_{j=1}^{r_{i\ell}} \widetilde{\sigma}_{i\ell j}^{-q_i}, \tag{11}$$

where $P_{i\ell} = V_{i\ell} V_{i\ell}^\top$ is the right-singular support introduced in Sec. 2.2, $\widetilde{\sigma}_{i\ell j} = \max(\sigma_{i\ell j}, \epsilon)$ is the stabilized singular value, and $q_i \geq 0$ controls the inverse spectral response. The singular values affect $C_{i\ell}$ only through scalar $s_{i\ell}$, leaving all retained directions isotropically weighted within $P_{i\ell}$.

Substituting Eq. (11) into the covariance-weighted solution in Eq. (3) gives

$$M_\ell^\star = \left(\sum_i w_i s_{i\ell} \Delta_{i\ell} P_{i\ell} \right) \left(\sum_i w_i s_{i\ell} P_{i\ell} + C_{0\ell} + \lambda I \right)^{-1}. \tag{12}$$

When $q_i = 0$, $s_{i\ell} = 1$ and the singular values do not affect the task operator. Positive q_i introduces an inverse spectral response, with larger exponents making the scalar more sensitive to smaller retained singular values. The choice of q_i must therefore determine the strength of this response without relying on task data.

3.2 Data-Free Calibration of the Spectral Exponent

The controlled analysis establishes that an inverse spectral response is preferable, but it does not determine how sharp this response should be. $q_i = 1$ gives the inverse mean $G_{i\ell}(1) = \frac{1}{r_{i\ell}} \sum_{j=1}^{r_{i\ell}} \widetilde{\sigma}_{i\ell j}^{-1}$, which can be highly sensitive to the smaller retained singular values. We therefore use a shared fractional exponent $q_i = q$, with $q \in [0, 1]$, and determine q directly from the checkpoint spectra.

To calibrate q without task data, define the aggregate spectral mass

$$F(q) = \sum_{i,\ell} r_{i\ell} G_{i\ell}(q) = \sum_{i,\ell,j} \widetilde{\sigma}_{i\ell j}^{-q}. \tag{13}$$

We use the average per-task mass of the sharp inverse response as the reference,

$$B = \frac{1}{K} F(1) = \frac{1}{K} \sum_{i,\ell,j} \widetilde{\sigma}_{i\ell j}^{-1}, \tag{14}$$

and choose the fractional exponent that most closely matches this reference:

$$q_{\text{free}}^\star = \arg\min_{q \in [0,1]} |F(q) - B|. \tag{15}$$

The resulting task operator uses $s_{i\ell} = G_{i\ell}(q_{\text{free}}^\star)$. The factor $1/K$ normalizes the reference spectral mass by the number of merged tasks. In the final merger, we therefore use $w_i = 1$ in the closed-form solution (12). The task-count normalization is incorporated through the calibrated spectral strength instead of an additional uniform task weight. This calibration is data-free since it uses only the singular values of the task checkpoints and requires no task examples.

3.3 Optional Classification-Aware Refinement

The shared calibration above treats all tasks with the same spectral exponent. For classification tasks, we optionally allow this exponent to vary mildly with the number of classes. The motivation is that the difficulty of preserving the largest competing logit among N_i classes grows with $\sqrt{\log N_i}$ under a standard sub-Gaussian bound; a formal statement is provided in Appendix A.1.

For task i with N_i classes, define $d_i = \sqrt{\log N_i}$ and $h_i = \frac{d_i}{\frac{1}{K}\sum_{k=1}^{K} d_k}$, so that the task factors have unit mean. We then set $q_i = qh_i$ and $s_{i\ell} = G_{i\ell}(qh_i) = \frac{1}{r_{i\ell}} \sum_{j=1}^{r_{i\ell}} \widetilde{\sigma}_{i\ell j}^{-qh_i}$. Thus, h_i changes how strongly task i responds to its inverse spectrum; it does not act as an additional task weight.

The shared scalar q is calibrated using the same checkpoint-only mass-matching rule, replacing $G_{i\ell}(q)$ with $G_{i\ell}(qh_i)$ while keeping the same task-normalized sharp-inverse reference. The classification-aware version is used as the default SOS-Merge configuration in our classification experiments, while the class-free variant remains applicable when class counts are unavailable.

Final merged model. After solving for $M_\ell^\star$, we form the merged parameter as $\theta_\ell^{\text{merge}} = \theta_{0\ell} + \alpha_\ell M_\ell^\star$, where $\alpha_\ell = 0.5 + \frac{b_\ell}{b_{\max}}$ is a fixed depth-dependent multiplier for visual residual blocks following (Wang et al., 2025) and $\alpha_\ell = 1$ elsewhere. Here b_ℓ is the zero-based index of the visual residual block containing parameter ℓ, and $b_{\max}$ is the largest residual-block index in the backbone. Appendix B.4 provides the algorithm and complexity analysis of the proposed method.

4 Experiments

4.1 Experimental Setup

Models and tasks. Following the CLIP model-merging protocol used in recent work (Xu et al., 2025; Marczak et al., 2025; Xu et al., 2026), we evaluate ViT-B/32, ViT-B/16, and ViT-L/14 pretrained CLIP vision encoders (Radford et al., 2021). For each backbone, we merge task-specific fine-tuned checkpoints over the 8-, 14-, and 20-task prefixes of a common 20-task pool, yielding 9 backbone–task-count settings. All methods within a setting use the same pretrained checkpoint, task-specific checkpoints, and test splits.

Evaluation protocol. We report the unweighted mean test top-1 accuracy over all tasks in each setting. Our main evaluation is strictly data-free: merging methods may use the pretrained and task-specific checkpoints, but no task examples, held-out accuracy, or validation-based model selection. Hyperparameters are fixed before test evaluation according to each method's official or recommended data-free configuration. Fine-tuned single-task expert models are reported only as an upper reference.

Implementation. Unless stated otherwise, SOS-Merge uses the classification-aware variant with the same configuration across all settings. We use rank 48, ridge $\lambda = 0.2$, and singular-value floor $\epsilon = 10^{-6}$. The floor does not alter any retained singular value in our experiments.

Details about task composition, datasets, baselines and implementation are provided in Appendix B.

4.2 Comparison with Recent Data-Free Mergers

We first evaluate whether the scale-only spectral design translates into a consistent advantage over existing data-free merging methods. In particular, we are interested in whether the benefit persists across different model scales and becomes more visible as more task-specific experts are combined. We therefore compare SOS-Merge with recent weight-space, subspace-based, and covariance-based model merging baselines.

Tab. 4 shows that SOS-Merge achieves the highest average accuracy, reaching 86.92% and improving over the strongest baseline, SWUDI-A, by 1.16 percentage points. It ranks first in eight of nine settings; the only exception is ViT-L/14 with eight tasks, where ISO-CTS+SVC is higher by 0.18 points. The advantage becomes larger as the number of merged tasks increases. On the 20-task benchmark, SOS-Merge improves over SWUDI-A by 2.20, 1.40, and 1.54 points on ViT-B/32,

Table 4: **Macro test top-1 accuracy (%) under the strictly data-free protocol.** Bold and underline denote the best and second-best merged methods, respectively. Performance of "Fine-tuned" expert models is reported as an upper reference. Complete per-task results and the few-shot setup are provided in Appendices C.1 and C.4.

	ViT-B/32			ViT-B/16			ViT-L/14			Average
Method	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	↑
Fine-tuned	90.53	89.20	89.99	92.61	91.21	91.94	94.14	92.55	93.28	91.72
Task Arithmetic	68.95	51.07	35.57	74.02	59.47	42.10	84.35	75.62	59.91	61.23
Average	66.02	64.35	60.94	71.71	69.33	65.20	79.55	76.76	71.57	69.49
TIES	72.57	63.94	57.36	76.74	70.50	64.75	85.44	78.65	74.15	71.57
ISO-C	82.74	78.74	73.60	87.14	83.37	77.90	91.15	88.02	86.84	83.28
ISO-CTS	81.07	79.84	76.85	86.60	84.78	81.49	90.68	88.95	88.67	84.33
TSV-M	83.88	79.36	75.93	87.23	83.83	79.92	91.34	88.24	87.18	84.10
WUDI	84.81	77.99	62.84	88.57	82.39	69.71	92.31	88.64	82.50	81.08
ACTMat	83.50	78.53	75.54	87.83	83.45	79.94	92.31	88.80	87.98	84.21
SWUDI-A	85.36	81.16	77.98	89.08	<u>85.42</u>	<u>81.78</u>	92.46	<u>89.91</u>	88.69	<u>85.76</u>
ACE-Merging	<u>85.41</u>	<u>81.24</u>	<u>78.04</u>	89.10	85.20	81.52	<u>92.54</u>	89.87	<u>88.78</u>	85.74
DC-Merge	80.57	78.33	76.54	84.66	82.66	79.67	89.48	87.59	86.84	82.93
ISO-CTS + SVC	85.15	80.32	76.63	<u>89.36</u>	84.95	81.35	**92.63**	89.19	88.54	85.35
SOS-MERGE	**86.39**	**83.05**	**80.18**	**89.40**	**86.65**	**83.17**	92.45	**90.73**	**90.23**	**86.92**

Table 5: **Comparison of scalar spectral statistics under matched layerwise total scale.** All variants use the same right-singular support and the same total scalar weight at each layer. The calibrated fractional inverse responses outperform the tested simple inverse-magnitude statistics, with the classification-aware variant achieving the highest average.

	ViT-B/32			ViT-B/16			ViT-L/14			Average
Scale on $P_{i\ell}$	8	14	20	8	14	20	8	14	20	↑
$1/\|\boldsymbol{\sigma}_{i\ell}\|_2$	86.28	82.62	79.69	89.38	86.36	82.81	92.38	90.43	89.71	86.63
$1/\|\boldsymbol{\sigma}_{i\ell}\|_1$	86.26	82.55	79.53	89.35	86.27	82.77	92.35	90.46	89.58	86.57
$1/\sigma_{i\ell 1}$	86.31	82.64	79.64	89.36	86.35	82.85	92.42	90.50	89.79	86.65
$G_{i\ell}(1)$	86.23	82.38	79.34	89.29	86.07	82.47	92.32	90.39	89.66	86.46
$G_{i\ell}(q^{\star}_{\text{free}})$	86.24	82.70	79.90	89.36	86.37	83.04	92.37	90.61	90.12	86.75
$G_{i\ell}(q^{\star}_{\text{cls}}h_i)$	**86.39**	**83.05**	**80.18**	**89.40**	**86.65**	**83.17**	**92.45**	**90.73**	**90.23**	**86.92**

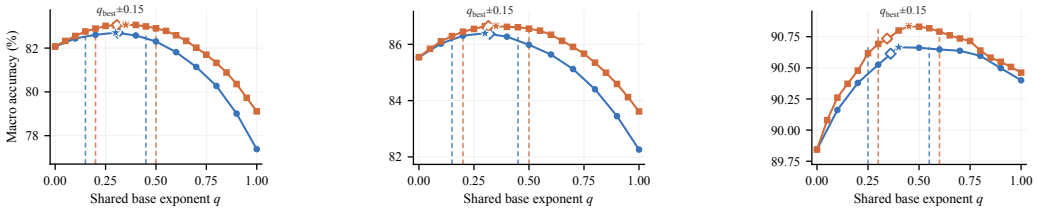
ViT-B/16, and ViT-L/14, respectively. Classical weight-space methods degrade more substantially as the task count grows, while recent subspace- and covariance-based methods remain stronger but still fall below SOS-MERGE in all three 20-task settings. These results show that our scale-only design translates into consistent gains across the three evaluated backbones and task counts, with the clearest separation from existing data-free methods when more tasks are merged.

4.3 COMPARING SCALAR SPECTRAL STATISTICS

Sec. 2 shows that singular values are more effective when summarized as a scalar spectral strength. This leaves a practical question: whether the fractional inverse moment used by SOS-MERGE is necessary, or whether a simpler inverse-magnitude statistic can provide an equally effective task allocation. We compare against the inverse Frobenius norm, inverse nuclear norm, and inverse leading singular value, together with the fixed and calibrated inverse moments used by SOS-MERGE.

A direct comparison of these statistics would also change the overall scale of the task operators. To isolate how each statistic allocates strength across tasks, we match their layerwise total scale to that of the classification-aware SOS-MERGE configuration. For a candidate statistic $f_{i\ell}$, we set $s^{(f)}_{i\ell} = \frac{f_{i\ell}}{\sum_{k=1}^{K} f_{k\ell}} \left(\sum_{k=1}^{K} G_{k\ell}(q^{\star}_{\text{cls}}h_k)\right)$. All variants therefore use the same right-singular support and the same total scalar weight at each layer. Only the allocation of this weight across tasks changes.

Tab. 5 shows that simple inverse-magnitude statistics already provide strong scalar weights. The best of these, $1/\sigma_{i\ell 1}$, reaches an average accuracy of 86.65%, compared with 86.75% for the class-free calibrated fractional moment. The classification-aware refinement further raises the average to 86.92% and improves over the simple statistics in all nine settings. The gains are modest, suggesting that the main benefit comes from using an appropriate scalar spectral strength, while fractional calibration and class-aware modulation provide additional refinements.



(a) ViT-B/32, $K = 14$. (b) ViT-B/16, $K = 14$. (c) ViT-L/14, $K = 14$.

Figure 2: **Diagnostic sweep of the spectral exponent q in the 14-task setting.** Curves show test accuracy for the class-free and classification-aware spectral responses as q varies, and markers indicate the exponents selected using checkpoint-only calibration. The average gap between the data-free calibration and the best sampled exponent is 0.05 for the scale-only setting and 0.06 for the classification-aware variant. The sweep is performed only for analysis and is not used for hyperparameter selection or the main results.

Table 6: **Ablation of auxiliary components.** Results are averaged over the 8-, 14-, and 20-task settings for each backbone. Removing the pretrained covariance or the post-solve depth multiplier reduces the performance of SOS-MERGE. Applying the same multiplier to ACE-Merging improves its performance, while SOS-MERGE remains higher under the matched setting.

Setting	ViT-B/32	ViT-B/16	ViT-L/14	Average
SOS-MERGE	**83.21**	**86.41**	**91.14**	**86.92**
w/o pretrained covariance	82.68	86.13	91.06	86.62
w/o depth multiplier	81.76	85.58	90.79	86.04
ACE-Merging	81.56	85.27	90.40	85.74
ACE-Merging + same depth multiplier	82.73	85.72	90.69	86.38

4.4 VALIDATING DATA-FREE EXPONENT CALIBRATION

The spectral exponent controls how strongly smaller retained singular values affect the scalar task weight. Although it could be selected using held-out accuracy, doing so would violate the data-free setting. We therefore examine whether the checkpoint-only calibration in Sec. 3.2 can achieve performance close to an exponent selected post hoc from an accuracy sweep. This sweep is used only for diagnosis and is never used by SOS-MERGE or in the main results.

Fig. 2 shows the accuracy as a function of q for the 14-task setting. We report both the class-free response $G_{i\ell}(q)$ and the classification-aware response $G_{i\ell}(qh_i)$, together with the exponents selected by their corresponding checkpoint-only calibration. In both cases, the calibrated configuration achieves accuracy close to the best sampled exponent across the three backbones.

The same comparison over all 9 backbone–task-count settings gives an average accuracy gap of only 0.05 percentage points between the class-free calibration and the best sampled exponent. The corresponding gap for the classification-aware variant is 0.06 points. Together with the comparison in Tab. 5, these results show that checkpoint-only mass matching avoids the sensitivity of the fixed $q = 1$ response while achieving performance close to post-hoc exponent selection, without using task data for model selection.

4.5 AUXILIARY ABLATIONS

Finally, we examine whether the gains of SOS-MERGE can be attributed to auxiliary components outside the scale-only spectral operator. We focus on two components used in the main configuration: the pretrained covariance term $C_{0\ell}$ and the fixed depth-dependent multiplier applied after solving the merger. We also apply the same depth multiplier to ACE-Merging to separate its effect from the difference between the two merging operators.

Tab. 6 shows that removing the pretrained covariance reduces the average accuracy from 86.92% to 86.62%. Removing the post-solve depth multiplier produces a larger drop to 86.04%. The multiplier is also beneficial to ACE-Merging, improving its average accuracy from 85.74% to 86.38%. Under the matched setting with the same multiplier, SOS-MERGE remains higher at 86.92%. These results indicate that both auxiliary components contribute to the final performance, with the depth-dependent multiplier having the larger effect. However, the advantage of SOS-MERGE is preserved

when the multiplier is controlled, suggesting that the improvement over ACE-Merging cannot be explained by this post-processing step alone.

Appendix C.2 provides additional analyses: a further decomposition of the scalar spectral weight, evaluations on random fixed-size task subsets, sensitivity studies of the retained rank and ridge coefficient, and results for methods that additionally support few-shot or validation-based selection.

5 Related Work

Weight-space and task-vector merging. Model merging commonly starts from task vectors, i.e., the parameter updates from a shared pretrained checkpoint to task-specific experts. Task Arithmetic (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023) uses the approximate local linearity of fine-tuned models to add such updates directly, while model soups and weight averaging (Wortsman et al., 2022) merge checkpoints in the same weight-space view. Later methods improve this coordinate-level operation by resolving signs, pruning unreliable magnitudes, or sparsifying task updates, as in TIES-Merging (Yadav et al., 2023) and Model Breadcrumbs (Davari & Belilovsky, 2024). Other work studies where task information concentrates and how task or layer scaling affects interference (Wang et al., 2024; 2025). Overall, these methods mainly manipulate task-vector coordinates, signs, magnitudes, or task-level weights.

Subspace and spectral merging. Recent data-free methods increasingly use matrix structure rather than treating updates as unstructured vectors. ISO-C and ISO-CTS (Marczak et al., 2025) decompose updates into common and task-specific subspaces to reduce interference, TSV-M (Gargiulo et al., 2025) merges through task singular-vector bases with low-rank structure, and projective merging (Wei et al., 2025) formulates multi-task merging through adaptive projections. DC-Merge (Zhang et al., 2026) uses directional consistency, and shared-knowledge correction (Li et al., 2026) further considers shared spectral components across tasks. SWUDI and SWUDI-A (Wei et al., 2026) regularize the spectrum of the closed-form merging system to control interference. Together, these methods demonstrate the value of subspace and spectral structure in data-free merging. Our analysis focuses on a different question: once the right-singular support and total spectral mass are controlled, is direction-specific singular-value weighting still beneficial?

Covariance and activation-inspired closed-form merging. Another line views merging as matching the behavior of task experts. Fisher merging (Matena & Raffel, 2022) uses an approximate local loss geometry, AdaMerging (Yang et al., 2024) learns merging coefficients from unlabeled validation examples, and ProDistill (Xu et al., 2025) performs progressive layer-wise activation matching with few-shot data. In the data-free setting, WUDI (Cheng et al., 2025) constructs checkpoint-derived surrogates for task-dependent activation information, while ACTMat (Hameed et al., 2026) estimates covariance information for closed-form merging directly from model parameters. ACE-Merging (Xu et al., 2026) brings this activation-matching perspective into the data-free setting by using task-update singular structure, in particular $\Delta^\top \Delta = V \Sigma^2 V^\top$, as a covariance proxy. Such covariance operators typically encode support and anisotropic spectral strength together. Our controlled analysis separates these roles and finds that, under matched support and total spectral mass, the full anisotropic spectrum performs worse than an isotropic operator whose spectrum is compressed into a single scalar strength.

6 Conclusion

We revisit the role of singular structure in data-free model merging by separating three coupled factors: task support, within-support anisotropy, and aggregate spectral strength. Our analysis shows that the right-singular subspace provides useful task support, while singular values are more effective when compressed into a scalar strength than when retained as direction-specific weights. Based on this finding, we propose SOS-Merge, a scale-only spectral merger that combines right-singular support with a checkpoint-calibrated fractional inverse response. The method requires no task examples or validation-based model selection and achieves strong performance across three CLIP backbones and 8-, 14-, and 20-task merging settings. These results suggest that, for data-free merging, separating where a task acts from how strongly it contributes can be more effective than preserving the full anisotropic spectral structure.

AI USE STATEMENT

We employed large language models (LLMs) in the preparation of this manuscript to assist with academic writing polishing, improve the overall clarity, fluency and readability of the full text, and identify and correct potential typographical errors. Additionally, LLMs were utilized for literature retrieval and academic resource discovery, helping us screen and sort relevant prior work to support the research background and literature review section of this study. All AI-generated and AI-polished content has been fully reviewed, verified, and revised manually by the authors. The manual inspection covers all polished textual paragraphs, the integrity and accuracy of every cited reference and literature resource, as well as all analytical content and code involved in the research. The authors take full and ultimate responsibility for all the content, viewpoints, data and conclusions presented in this manuscript.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide the code, dataset and model in the supplementary materials and anonymous github `https://anonymous.4open.science/r/Singular-Structure-in-Data-Free-Model-Merging-Revisited-1EA6`. In the README.md file included with the code, we present a step-by-step guide for reproducing our results.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=xR9msNaREW`.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018. URL `http://arxiv.org/abs/1812.01718`. cite arxiv:1812.01718Comment: To appear at Neural Information Processing Systems 2018 Workshop on Machine Learning for Creativity and Design.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `https://proceedings.mlr.press/v15/coates11a.html`.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scalable upcycling of finetuned foundation models via sparse task vectors merging. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL `https://openreview.net/forum?id=vuyP3tupig`.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025. doi: 10.1109/CVPR52734.2025.01742.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2014.09.005. URL https://www.sciencedirect.com/science/article/pii/S0893608014002159. Special Issue on "Deep Learning of Representations".

Marawan Gamal Abdel Hameed, Derek Tam, Pascal Jr Tikeng Notsawo, Colin Raffel, and Guillaume Rabusseau. Model merging via data-free covariance estimation, 2026. URL https://arxiv.org/abs/2604.01329.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yayuan Li, Ze Peng, Jian Zhang, Jintao Guo, Yue Duan, and Yinghuan Shi. When shared knowledge hurts: Spectral over-accumulation in model merging. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WUjk3RVZyV.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=RBZpAa27ls.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=LSKlp_aceOC.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=0A9f2jZDGW.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2012.02.016. URL https://www.sciencedirect.com/science/article/pii/S0893608012000457. Selected Papers from IJCNN 2011.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pp. 210–218, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-00933-5. doi: 10.1007/978-3-030-00934-2_24. URL https://doi.org/10.1007/978-3-030-00934-2_24.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *International Conference on Machine Learning*, 2024.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. Lines: Post-training layer scaling prevents forgetting and enhances model merging. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=J5sUOvlLbQ.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66178–66193. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wei25k.html.

Yongxian Wei, Runxi Cheng, Xingxuan Zhang, Li Shen, Chun Yuan, Peng Cui, and Dacheng Tao. Closed-form spectral regularization for multi-task model merging, 2026. URL https://arxiv.org/abs/2606.07289.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv e-prints*, art. arXiv:1708.07747, August 2017. doi: 10.48550/arXiv.1708.07747.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Bo Xu, Haotian Wu, Hehai Lin, Weiquan Huang, Beier Zhu, Yao Shu, and Chengwei Qin. Acemerging: Data-free model merging with adaptive covariance estimation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 29472–29482, June 2026.

Jing Xu, Jiazheng Li, and Jingzhao Zhang. Scalable model merging with progressive layer-wise distillation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=xX8NJShgny.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nZP6NgD3QY.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Comput. Surv.*, 58(8), February 2026. ISSN 0360-0300. doi: 10.1145/3787849. URL https://doi.org/10.1145/3787849.

Han-Chen Zhang, Zi-Hao Zhou, Mao-Lin Luo, Shimin Di, Min-Ling Zhang, and Tong Wei. Dc-merge: Improving model merging with directional consistency. In *CVPR*, 2026.

A Theoretical Motivation for the Classification-Aware Refinement

A.1 Competing-Class Effect

The classification-aware variant in Sec. 3.3 mildly adjusts the spectral exponent according to the number of classes in each task. The motivation is that top-1 classification depends on maintaining the logit margin against every competing class. As the number of competing classes increases, the largest perturbation among these margins can also increase, even when the perturbation associated with each individual class has the same scale. The following result makes this dependence explicit under a standard sub-Gaussian assumption.

Assumption 1 (Competing logit-gap perturbations)**.** *For a sample from task i with label y and N_i classes, let z_i be the task expert logits and let δz_i be the logit perturbation introduced by merging. For each competing class $c \neq y$, define the centered logit-gap perturbation*

$$\xi_{i,c} = \big(\delta z_{i,c} - \delta z_{i,y}\big) - \mathbb{E}\big[\delta z_{i,c} - \delta z_{i,y}\big]. \tag{16}$$

We assume each $\xi_{i,c}$ is sub-Gaussian with variance proxy at most ν_i^2. No independence across competing classes is required.

Theorem 1 (Competing-class maximum)**.** *Under Assumption 1, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{c \neq y} |\xi_{i,c}| \leq \nu_i \sqrt{2 \log \frac{2(N_i - 1)}{\delta}}. \tag{17}$$

For any competing class $c \neq y$, Assumption 1 implies the two-sided sub-Gaussian tail bound

$$\Pr\left(|\xi_{i,c}| > t\right) \leq 2 \exp\left(-\frac{t^2}{2\nu_i^2}\right). \tag{18}$$

Applying the union bound over the $N_i - 1$ competing classes gives

$$\Pr\left(\max_{c \neq y} |\xi_{i,c}| > t\right) \leq 2(N_i - 1) \exp\left(-\frac{t^2}{2\nu_i^2}\right). \tag{19}$$

Setting the right-hand side to δ yields

$$t = \nu_i \sqrt{2 \log \frac{2(N_i - 1)}{\delta}}, \tag{20}$$

which proves the stated high-probability bound.

For a fixed confidence level δ, the bound grows as $O(\sqrt{\log N_i})$. Thus, even when individual competing-class perturbations have comparable scale, tasks with more classes can experience a larger worst-case logit-gap perturbation. This motivates the class-count factor used in Sec. 3.3, Note that the result does not imply that $q_i = qh_i$ is an optimal exponent rule.

B Experimental Setup and Implementation

B.1 Datasets

We follow the CLIP vision model-merging benchmark used by Representation Surgery (Wang et al., 2024), ISO-C/ISO-CTS (Marczak et al., 2025), TSV-M (Gargiulo et al., 2025), and ACE-Merging (Xu et al., 2026). We evaluate nested settings with 8, 14, and 20 tasks. The 8-task setting consists of Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), GTSRB (Stallkamp et al., 2012), MNIST (Lecun et al., 1998), RESISC45 (Cheng et al., 2017), SVHN (Netzer et al., 2011), and SUN397 (Xiao et al., 2010).

The 14-task setting extends the 8-task setting with STL10 (Coates et al., 2011), OxfordIIIT-Pet (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), CIFAR100 (Krizhevsky, 2009), PCAM (Veeling et al., 2018), and FER2013 (Goodfellow et al., 2015). The 20-task setting further adds CIFAR10 (Krizhevsky, 2009), Food101 (Bossard et al., 2014), FashionMNIST (Xiao et al., 2017), RenderedSST2 (Radford et al., 2021), EMNIST (Cohen et al., 2017), and KMNIST (Clanuwat et al., 2018).

For each setting, we evaluate the merged model separately on every included dataset and report the unweighted mean of test top-1 accuracy across tasks.

B.2 Baselines

We compare SOS-Merge with the following data-free model-merging baselines. Variants that use task examples, few-shot validation data, or held-out accuracy are reported separately in Appendix C.4; Tab. 4 includes only strictly data-free configurations.

- **Task Arithmetic** (Ilharco et al., 2023) constructs task vectors as parameter differences from the pretrained checkpoint and merges them through weighted addition.
- **Average** (Wortsman et al., 2022) directly averages the task-specific checkpoints in weight space.
- **TIES** (Yadav et al., 2023) trims small-magnitude update coordinates, resolves sign conflicts across task vectors, and merges the retained aligned coordinates.
- **ISO-C** (Marczak et al., 2025) applies SVD to the aggregated task update and replaces its singular-value spectrum with an isotropic scale.
- **ISO-CTS** (Marczak et al., 2025; Gargiulo et al., 2025) extends ISO-C by combining a common singular subspace with task-specific singular directions.
- **TSV-M** (Gargiulo et al., 2025) merges selected task singular vectors and reconstructs the update using isotropic scaling.
- **WUDI** (Cheng et al., 2025) performs data-free activation matching using surrogate inputs constructed from task-update information.
- **ACTMat** (Hameed et al., 2026) estimates covariance information without task data and incorporates it into a closed-form merging system.
- **SWUDI-A** (Wei et al., 2026) applies adaptive spectral regularization to the closed-form merging system.
- **ACE-Merging** (Xu et al., 2026) constructs covariance-weighted merging operators from the singular structure of task updates.
- **DC-Merge** (Zhang et al., 2026) selects sparse and decorrelated components of task updates before merging.

- **ISO-CTS+SVC** (Li et al., 2026) applies shared-vector correction to ISO-CTS.

For Tab. 4, all baseline configurations are fixed before test evaluation, without using task examples, few-shot validation data, held-out accuracy, or test accuracy for model selection. We use coefficient 0.3 for Task Arithmetic; coefficient 0.3 and density 0.2 for TIES; coefficient 1.0 for ISO-C, ISO-CTS, TSV-M, and DC-Merge; common-subspace fraction 0.8 for ISO-CTS; keep fraction 10^{-3} for DC-Merge; 300 optimization steps with learning rate 10^{-5} for WUDI; ACTMat uses coefficient 1.0 and the authors' default split-QKV ViT protocol. SWUDI-A uses coefficient 1.0, split QKV, and the participation-square-root adaptive-rank rule. $\epsilon = 10^{-5}$, γ threshold 0.3, and refinement-rank fraction 0.3 for ACE-Merging; and the default $1/K$ coefficient for SVC in ISO-CTS+SVC.

Single-task expert models are evaluated by applying each independently fine-tuned checkpoint to its own task and averaging the resulting test accuracies over the same task set. They are reported only as an upper reference and are neither merging baselines nor used for hyperparameter selection.

B.3 Implementation Details

All experiments use the same pretrained CLIP visual encoder and task-specific fine-tuned checkpoints within each backbone–task-count setting. During evaluation, we attach the corresponding zero-shot classification head for each dataset and evaluate the merged visual encoder without further training or adaptation.

SOS-Merge applies the covariance-weighted solver to eligible matrix parameters in the visual encoder. A parameter is eligible if it is two-dimensional, has both dimensions larger than one, is not `text_projection`, and is not an embedding parameter. For all remaining floating-point tensors, including biases and excluded parameters, we use the uniform task-update average $K^{-1}\sum_i \Delta_{i\ell}$. Thus, the spectral construction is applied only to matrix updates for which the right-singular support $V_{i\ell}V_{i\ell}^\top$ is defined.

For each eligible matrix, we retain rank $r_\ell = \min(48, d_\ell^{\text{out}}, d_\ell^{\text{in}})$ and construct the task operator using Eq. (11). Negative powers are evaluated using $\widetilde{\sigma} = \max(\sigma, \epsilon)$ with $\epsilon = 10^{-6}$. This floor does not modify any retained singular value in our experiments. The pretrained operator $C_{0\ell}$ is fixed to the retained input-side Gram of the pretrained matrix. For $W_{0\ell} \approx U_{0\ell}\Sigma_{0\ell}V_{0\ell}^\top$, we use $C_{0\ell} = V_{0\ell}\,\mathrm{diag}\big(\widetilde{\sigma}_{0\ell 1}^2, \ldots, \widetilde{\sigma}_{0\ell r_\ell}^2\big)\,V_{0\ell}^\top$, which corresponds to $W_{0\ell}^\top W_{0\ell}$ restricted to the retained right-singular subspace. We use $\lambda = 0.2$ for the ridge term in all main experiments.

The scalar calibration in Sec. 3.2 is performed over $q \in [0, 1]$ using only the cached task spectra. Matrix systems are assembled and solved in float64, after which the merged update is converted back to the checkpoint dtype. Following the closed-form solve, we apply a fixed depth-dependent multiplier to parameters inside visual residual blocks, $\alpha_\ell = 0.5 + \frac{b_\ell}{b_{\max}}$ where b_ℓ is the zero-based residual-block index and $b_{\max}$ is the largest block index of the backbone. Parameters outside residual blocks use $\alpha_\ell = 1$. The final parameter is $\theta_\ell^{\text{merge}} = \theta_{0\ell} + \alpha_\ell M_\ell^\star$.

B.4 Algorithm and Complexity

Complexity. Let an eligible matrix parameter have shape $m_\ell \times n_\ell$, where $m_\ell = d_\ell^{\text{out}}$ and $n_\ell = d_\ell^{\text{in}}$, and let $r_\ell = \min(r, m_\ell, n_\ell)$ be the retained rank. Computing rank-r_ℓ SVD factors for the K task updates and the pretrained matrix costs

$$O((K+1)m_\ell n_\ell r_\ell)$$

with a truncated SVD routine, or $O((K+1)m_\ell n_\ell \min\{m_\ell, n_\ell\})$ with a dense exact SVD. This is a one-time preprocessing cost, and the resulting factors are reused during calibration and merging.

Given the cached factors, the numerator term for each task can be assembled as $U_{i\ell}\,\mathrm{diag}(\sigma_{i\ell}s_{i\ell})V_{i\ell}^\top$ in $O(m_\ell n_\ell r_\ell)$ time, while $s_{i\ell}V_{i\ell}V_{i\ell}^\top$ costs $O(n_\ell^2 r_\ell)$. Constructing the pretrained operator contributes another $O(n_\ell^2 r_\ell)$. Solving the dense system in Eq. (12), whose coefficient matrix is $n_\ell \times n_\ell$ with m_ℓ right-hand sides, costs $O(n_\ell^3 + m_\ell n_\ell^2)$. The post-SVD time for layer ℓ is therefore

$$O\big(Kr_\ell(m_\ell n_\ell + n_\ell^2) + n_\ell^2 r_\ell + n_\ell^3 + m_\ell n_\ell^2\big)\,.$$

The total merge cost is obtained by summing this expression over eligible matrix parameters $\ell \in \mathcal{L}$, plus the SVD preprocessing cost when the factors are not already cached.

Algorithm 1 Scale-only spectral merging

Require: Pretrained checkpoint θ_0; task checkpoints $\{\theta_i\}_{i=1}^K$; task class counts $\{N_i\}_{i=1}^K$ for the classification-aware variant; rank r; ridge λ; singular-value floor ϵ.

Form task updates $\Delta_i \leftarrow \theta_i - \theta_0$
for each eligible matrix parameter ℓ **do**
  Set $r_\ell \leftarrow \min(r, d_\ell^{\text{out}}, d_\ell^{\text{in}})$
  Compute rank-r_ℓ SVDs $\Delta_{i\ell} \approx U_{i\ell}\Sigma_{i\ell}V_{i\ell}^\top$ for all tasks
  Set $P_{i\ell} \leftarrow V_{i\ell}V_{i\ell}^\top$ and $\widetilde{\sigma}_{i\ell j} \leftarrow \max(\sigma_{i\ell j}, \epsilon)$
  Compute the fixed pretrained operator $C_{0\ell}$ from the rank-r_ℓ SVD of $W_{0\ell}$
end for
if using the class-free variant **then**
  Set $h_i \leftarrow 1$ for all tasks
else
  Set $h_i \leftarrow \sqrt{\log N_i}/(K^{-1}\sum_{k=1}^K \sqrt{\log N_k})$
end if
Choose

$$q^\star \leftarrow \arg\min_{q\in[0,1]} \left| \sum_{i,\ell,j} \widetilde{\sigma}_{i\ell j}^{-qh_i} - \frac{1}{K}\sum_{i,\ell,j} \widetilde{\sigma}_{i\ell j}^{-1} \right|$$

for each parameter tensor ℓ **do**
  if ℓ is an eligible matrix parameter **then**
    Set $s_{i\ell} \leftarrow r_\ell^{-1}\sum_{j=1}^{r_\ell} \widetilde{\sigma}_{i\ell j}^{-q^\star h_i}$
    Compute $M_\ell^\star$ from Eq. (12) with $w_i = 1$
  else
    $M_\ell^\star \leftarrow K^{-1}\sum_i \Delta_{i\ell}$
  end if
  Set $\alpha_\ell \leftarrow 0.5 + b_\ell/b_{\max}$ inside a residual block and $\alpha_\ell \leftarrow 1$ otherwise
  $\theta_\ell^{\text{merge}} \leftarrow \theta_{0\ell} + \alpha_\ell M_\ell^\star$
end for
return merged checkpoint θ^{merge}

Exponent calibration is lower order because each objective evaluation involves only scalar operations over the retained singular values. With B evaluations of the one-dimensional calibration objective, its time complexity is $O\big(BK\sum_{\ell\in\mathcal{L}} r_\ell\big)$, with $O(1)$ additional working memory beyond the cached spectra. The classification-aware variant has the same asymptotic complexity because it only replaces $q^\star$ by $q^\star h_i$ in the scalar spectral response and requires no additional SVDs, forward passes, or accuracy evaluations.

For one matrix layer, the dense numerator and denominator require $O(m_\ell n_\ell + n_\ell^2)$ working memory. Cached SVD factors require $O((K+1)r_\ell(m_\ell+n_\ell))$ storage per layer. If factors for all layers are cached, their total storage is $O\big((K+1)\sum_{\ell\in\mathcal{L}} r_\ell(m_\ell+n_\ell)\big)$, and the peak merge-time working memory is $O(\max_{\ell\in\mathcal{L}}\{m_\ell n_\ell + n_\ell^2\})$ in addition to the checkpoint and cached factors.

C Additional Experiment Results

C.1 Complete Results for the Main Experiments

Tab. 4 reports the macro test accuracy for each backbone–task-count setting. Here we provide the corresponding per-task results for all nine settings (Tab. 7, Tab. 8, Tab. 9, Tab. 10, Tab. 11, Tab. 12, Tab. 13, Tab. 14, Tab. 15). All methods follow the same evaluation protocol as in the main experiments, and the final row of each table reports the unweighted mean accuracy across the included tasks.

Table 7: **Per-task test accuracy (%) for ViT-B/32 with 8 tasks.** The final row reports the unweighted mean over the eight tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	54.94	63.34	58.60	74.13	74.47	70.74	71.16	69.28	70.59	71.71	74.85	75.08	72.69	77.66
DTD	50.11	50.11	54.73	67.93	66.97	69.15	69.31	69.10	70.69	70.90	65.69	72.23	74.31	79.41
EuroSAT	77.41	73.63	81.81	92.78	90.41	94.63	95.33	93.48	96.93	95.89	92.96	95.22	97.44	99.89
GTSRB	69.68	52.80	70.86	85.98	83.33	91.94	94.51	92.72	94.57	92.97	82.72	91.59	95.90	98.73
MNIST	97.34	87.46	98.36	98.18	97.38	99.25	99.47	99.41	99.48	99.41	96.71	98.93	99.44	99.69
RESISC45	66.67	71.44	70.97	87.90	87.17	85.48	85.32	82.70	87.25	87.00	85.79	90.13	90.06	96.11
SVHN	80.21	64.16	86.15	83.07	76.82	92.02	94.27	92.81	93.85	94.06	74.99	86.12	92.48	97.46
SUN397	55.27	65.25	59.11	71.95	72.04	67.83	69.07	68.51	69.54	71.36	70.84	71.89	68.80	75.28
Average	68.95	66.02	72.57	82.74	81.07	83.88	84.81	83.50	85.36	85.41	80.57	85.15	86.39	90.53

Table 8: **Per-task test accuracy (%) for ViT-B/32 with 14 tasks.** The final row reports the unweighted mean over the 14 tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	27.73	60.71	46.14	60.70	68.06	57.59	59.12	63.01	63.95	67.94	67.52	63.95	65.94	77.66
DTD	36.86	47.66	49.10	64.68	66.44	64.73	61.22	64.47	65.80	66.44	61.81	68.56	71.17	79.41
EuroSAT	56.15	63.70	67.67	89.74	89.48	92.22	90.07	82.93	93.37	90.63	89.33	90.37	94.70	99.89
GTSRB	39.52	43.89	55.23	80.94	80.84	85.18	85.94	85.87	89.10	86.49	76.52	83.67	89.76	98.73
MNIST	86.56	76.43	93.80	98.02	97.79	98.94	98.99	99.08	99.22	99.24	95.31	98.53	99.23	99.69
RESISC45	44.52	66.87	62.57	82.37	84.60	80.38	73.73	77.43	81.68	81.78	82.78	85.25	86.75	96.11
SVHN	45.28	49.39	60.83	79.40	76.54	85.90	87.19	87.83	88.74	90.64	69.20	79.80	86.72	97.46
SUN397	36.60	64.50	55.86	68.20	69.86	65.30	62.83	67.24	68.03	70.26	69.06	68.47	68.24	75.28
STL10	83.69	97.11	91.07	96.99	97.85	97.15	96.08	96.40	97.31	97.56	98.24	97.38	97.92	97.65
OxfordIIITPet	69.80	88.09	80.29	89.48	90.71	89.86	88.42	88.50	90.19	90.27	91.77	89.02	90.87	90.65
Flowers102	41.49	68.29	55.78	78.06	81.44	75.64	74.03	71.78	78.66	73.49	79.61	81.27	84.68	88.70
CIFAR100	42.41	69.76	58.32	75.29	77.08	70.07	67.65	67.00	70.94	72.03	75.41	75.80	75.95	88.02
PCAM	61.60	62.63	72.64	82.15	81.03	84.37	83.60	81.93	83.90	84.19	79.44	82.27	83.11	87.00
FER2013	42.73	41.85	45.83	56.32	55.98	63.76	62.98	65.94	65.37	66.47	60.56	60.11	67.67	72.58
Average	51.07	64.35	63.94	78.74	79.84	79.36	77.99	78.53	81.16	81.24	78.33	80.32	83.05	89.20

Table 9: **Per-task test accuracy (%) for ViT-B/32 with 20 tasks.** The final row reports the unweighted mean over the 20 tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	8.73	59.43	35.07	38.76	57.32	45.24	32.12	54.48	53.64	59.86	63.91	51.21	53.67	77.66
DTD	22.45	46.06	41.76	58.78	62.93	61.06	46.22	60.80	61.65	62.07	61.22	64.68	68.83	79.41
EuroSAT	34.96	57.07	49.48	80.15	84.15	87.07	65.41	77.93	89.41	85.85	87.70	84.70	91.63	99.89
GTSRB	20.46	41.40	42.36	72.22	75.87	75.75	56.98	77.28	81.22	76.91	70.93	74.96	82.83	98.73
MNIST	64.48	70.48	84.48	95.53	96.01	94.53	91.44	92.17	94.77	92.57	88.24	96.72	95.82	99.69
RESISC45	19.19	64.43	53.05	73.38	79.95	73.19	51.16	72.86	77.56	77.90	80.11	79.32	84.22	96.11
SVHN	31.71	47.31	53.79	76.46	73.86	77.67	61.64	78.43	81.76	83.82	66.46	73.89	77.97	97.46
SUN397	16.27	64.00	52.52	62.57	66.62	62.86	45.21	65.67	66.15	68.56	68.26	64.81	67.45	75.28
STL10	68.65	96.85	89.29	94.97	97.20	96.14	90.34	95.91	96.62	96.70	98.00	96.67	97.76	97.65
OxfordIIITPet	41.59	87.74	75.74	84.98	89.37	87.68	76.51	87.14	88.99	88.80	91.61	87.44	90.05	90.65
Flowers102	19.63	66.74	49.78	72.35	78.00	70.68	54.37	68.27	75.43	70.29	78.34	76.66	81.40	88.70
CIFAR100	25.65	69.18	52.49	69.95	73.80	66.13	54.29	64.81	67.71	68.29	73.19	73.16	73.80	88.02
PCAM	51.36	62.38	69.67	76.21	82.22	84.15	77.91	81.81	83.07	83.83	81.08	80.29	82.63	87.00
FER2013	36.08	40.97	42.95	53.00	55.48	58.64	49.82	61.41	61.40	63.11	58.55	57.58	64.71	72.58
CIFAR10	62.12	92.67	86.42	93.62	95.15	93.21	88.19	92.22	93.66	93.28	94.60	94.64	95.06	97.77
Food101	12.10	79.26	48.87	68.74	76.53	72.03	43.16	72.40	78.30	78.82	83.07	72.91	81.73	85.59
FashionMNIST	67.96	70.67	73.75	83.17	84.37	86.04	78.98	83.97	85.64	84.07	83.15	84.47	86.64	95.26
RenderedSST2	50.85	60.41	56.07	69.25	69.74	70.13	54.26	70.84	71.72	72.76	71.77	69.52	70.79	74.03
EMNIST	39.92	33.07	66.95	91.20	86.55	97.40	95.26	98.53	97.46	98.47	87.77	90.35	96.27	99.83
KMNIST	17.22	8.68	22.65	56.69	51.88	59.08	43.55	53.96	53.44	54.77	42.86	58.60	60.38	98.43
Average	35.57	60.94	57.36	73.60	76.85	75.93	62.84	75.54	77.98	78.04	76.54	76.63	80.18	89.99

C.2 FURTHER DECOMPOSITION OF THE SCALAR SPECTRAL WEIGHT

The scale-only operator assigns one scalar spectral weight to each task and layer. We further examine whether this scalar mainly reflects the overall magnitude of the retained spectrum or also benefits

Table 10: **Per-task test accuracy (%) for ViT-B/16 with 8 tasks.** The final row reports the unweighted mean over the eight tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	65.97	69.79	70.43	82.30	83.07	80.40	81.68	79.54	81.72	81.79	82.81	83.66	82.33	86.79
DTD	52.71	50.05	57.13	71.91	71.17	72.66	73.62	75.64	75.37	76.91	68.72	77.18	79.31	82.07
EuroSAT	76.78	79.26	81.70	96.48	96.30	97.04	98.52	96.59	98.67	98.00	95.93	98.37	97.96	99.85
GTSRB	73.92	60.35	70.26	91.85	91.31	93.94	96.55	91.94	96.71	93.60	86.13	96.60	97.07	99.17
MNIST	98.48	94.45	98.80	98.80	98.38	99.27	99.44	99.46	99.47	99.47	97.98	99.13	99.41	99.76
RESISC45	74.49	75.98	79.79	91.92	92.03	89.10	90.13	91.03	91.21	91.70	89.67	93.68	92.37	96.89
SVHN	88.06	75.12	89.89	88.25	84.89	93.93	95.30	94.96	95.55	95.94	82.33	90.63	93.83	97.86
SUN397	61.75	68.64	65.89	75.57	75.62	71.54	73.30	73.49	73.96	75.38	73.75	75.61	72.90	78.51
Average	74.02	71.71	76.74	87.14	86.60	87.23	88.57	87.83	89.08	89.10	84.66	89.36	89.40	92.61

Table 11: **Per-task test accuracy (%) for ViT-B/16 with 14 tasks.** The final row reports the unweighted mean over the 14 tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	43.29	67.21	62.94	71.81	78.50	72.64	71.74	73.80	76.57	77.74	78.14	75.39	79.85	86.79
DTD	37.50	48.35	49.52	67.18	69.20	66.44	63.46	68.56	67.98	69.73	64.68	70.90	74.63	82.07
EuroSAT	55.63	68.70	74.26	95.81	96.67	95.56	94.74	93.56	96.70	95.70	94.74	96.96	96.56	99.85
GTSRB	58.07	53.44	62.03	89.01	90.89	90.86	89.18	87.13	93.96	88.30	82.88	93.40	94.38	99.17
MNIST	95.25	87.77	97.55	98.79	98.77	99.10	99.29	99.32	99.41	99.32	97.57	99.06	99.26	99.76
RESISC45	52.05	71.90	69.73	89.60	91.33	86.33	84.79	88.46	88.21	89.10	88.21	91.05	91.10	96.89
SVHN	73.02	65.74	81.22	87.95	86.33	91.18	90.55	93.35	93.26	93.98	79.47	88.81	89.31	97.86
SUN397	46.48	67.60	63.64	71.91	73.43	69.71	68.53	71.19	72.11	73.68	71.74	72.17	72.20	78.51
STL10	91.12	98.24	95.78	98.02	98.61	98.02	97.86	97.11	98.28	98.17	98.69	98.16	98.50	98.42
OxfordIIITPet	83.37	90.62	89.48	93.46	93.79	93.89	93.73	93.70	94.25	94.19	94.82	93.32	94.58	93.87
Flowers102	48.90	72.73	65.15	84.42	87.33	81.90	78.55	78.24	85.18	79.87	83.07	87.69	88.58	91.15
CIFAR100	48.73	74.46	65.77	80.49	81.84	76.20	74.88	75.80	77.44	78.61	79.32	80.89	79.31	89.46
PCAM	56.85	60.25	62.08	82.27	82.99	86.34	85.66	82.54	84.73	85.87	83.15	80.39	85.02	89.74
FER2013	42.31	43.61	47.85	56.48	57.29	65.42	60.43	65.46	67.80	68.54	60.74	61.09	69.81	73.38
Average	59.47	69.33	70.50	83.37	84.78	83.83	82.39	83.45	85.42	85.20	82.66	84.95	86.65	91.21

Table 12: **Per-task test accuracy (%) for ViT-B/16 with 20 tasks.** The final row reports the unweighted mean over the 20 tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	15.89	65.73	48.58	47.47	65.96	59.83	47.77	66.50	68.26	71.72	73.67	59.06	72.55	86.79
DTD	24.31	46.91	42.82	61.49	65.74	62.23	49.52	64.73	63.62	64.95	62.55	66.86	71.49	82.07
EuroSAT	35.33	62.07	63.67	91.30	95.74	92.59	76.30	90.56	95.00	93.33	94.22	94.78	94.89	99.85
GTSRB	35.62	51.26	53.75	83.51	88.24	83.54	67.47	77.93	88.35	79.71	78.37	88.44	89.02	99.17
MNIST	86.55	84.87	95.15	97.93	97.83	97.63	96.48	96.77	97.95	96.50	94.46	98.20	97.97	99.76
RESISC45	23.65	70.13	61.13	84.14	88.67	80.90	68.65	85.56	85.73	86.54	86.11	87.67	89.75	96.89
SVHN	52.93	63.97	74.69	86.11	85.13	87.70	79.16	88.02	88.91	88.97	76.05	86.46	84.67	97.86
SUN397	25.68	67.24	61.91	67.45	71.23	68.02	56.26	70.22	71.01	72.50	71.05	69.56	71.84	78.51
STL10	76.09	98.10	93.77	96.55	98.05	97.20	94.41	96.49	97.89	97.61	98.52	97.64	98.36	98.42
OxfordIIITPet	53.72	89.97	88.33	91.41	92.80	93.54	88.77	92.18	93.76	93.02	94.69	91.91	93.98	93.87
Flowers102	19.60	72.14	59.39	77.39	85.70	77.90	64.51	74.42	81.64	77.02	81.53	84.26	85.33	91.15
CIFAR100	29.86	74.21	61.40	74.41	78.70	72.07	62.51	72.27	74.95	75.77	77.46	77.60	78.12	89.46
PCAM	60.40	60.31	61.51	77.10	79.68	85.38	79.27	84.40	85.84	85.62	80.67	79.47	83.17	89.74
FER2013	37.32	40.99	43.98	54.28	56.12	60.03	43.31	60.96	63.07	63.99	58.33	57.48	65.81	73.38
CIFAR10	63.53	94.22	89.71	95.16	96.20	95.24	90.57	94.14	95.87	95.25	95.84	95.90	96.14	98.55
Food101	19.94	85.56	69.75	79.90	85.04	82.42	66.92	84.41	85.97	87.00	88.42	82.66	87.73	89.52
FashionMNIST	59.86	73.38	76.67	87.66	89.10	88.35	81.34	87.55	87.98	87.66	86.94	89.52	87.99	95.63
RenderedSST2	60.96	60.85	59.25	72.76	76.11	75.29	61.56	76.33	76.61	77.87	75.01	77.48	74.63	79.63
EMNIST	45.29	31.56	69.43	89.89	86.87	96.28	92.07	97.72	96.03	97.94	86.28	90.12	95.15	99.83
KMNIST	15.43	10.61	20.10	42.09	46.81	42.29	27.38	37.71	37.08	37.42	33.24	51.91	44.88	98.67
Average	42.10	65.20	64.75	77.90	81.49	79.92	69.71	79.94	81.78	81.52	79.67	81.35	83.17	91.94

from its normalized shape. Let $\bar{\sigma}_{i\ell} = r_{i\ell}^{-1} \sum_j \widetilde{\sigma}_{i\ell j}$. The calibrated scalar can be decomposed as

$$G_{i\ell}(q_i) = \bar{\sigma}_{i\ell}^{-q_i} \left(\frac{1}{r_{i\ell}} \sum_{j=1}^{r_{i\ell}} \left(\frac{\widetilde{\sigma}_{i\ell j}}{\bar{\sigma}_{i\ell}} \right)^{-q_i} \right) \equiv M_{i\ell} H_{i\ell}, \tag{21}$$

Table 13: **Per-task test accuracy (%) for ViT-L/14 with 8 tasks.** The final row reports the unweighted mean over the eight tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	82.13	81.56	84.18	90.92	91.16	89.85	91.01	90.54	91.15	91.02	90.47	91.88	90.71	92.39
DTD	65.64	62.77	67.93	79.15	77.87	79.10	81.17	81.01	81.22	81.60	75.96	82.61	82.18	84.15
EuroSAT	92.63	91.44	94.59	98.30	98.26	98.52	98.81	98.44	98.96	98.93	98.00	98.59	99.07	99.89
GTSRB	86.78	70.65	82.99	95.48	95.30	96.16	98.04	97.12	98.16	97.13	93.72	97.79	98.27	99.24
MNIST	98.94	97.01	99.00	99.24	99.18	99.51	99.63	99.67	99.65	99.62	99.06	99.42	99.68	99.69
RESISC45	86.65	82.60	88.89	94.81	94.70	93.98	94.00	95.03	94.22	95.22	93.87	96.03	94.94	97.37
SVHN	87.90	78.23	90.03	91.37	89.36	95.34	96.21	96.31	96.56	96.52	87.48	93.78	95.45	98.11
SUN397	74.13	72.14	75.90	79.93	79.62	78.28	79.59	80.38	79.77	80.30	77.29	80.95	79.29	82.32
Average	84.35	79.55	85.44	91.15	90.68	91.34	92.31	92.31	92.46	92.54	89.48	92.63	92.45	94.14

Table 14: **Per-task test accuracy (%) for ViT-L/14 with 14 tasks.** The final row reports the unweighted mean over the 14 tasks.

Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	67.34	79.79	75.56	86.56	89.39	85.55	87.39	87.69	88.27	88.37	88.20	89.23	89.57	92.39
DTD	56.17	59.41	60.43	77.13	78.24	75.59	75.59	77.82	77.98	78.30	72.18	79.26	80.64	84.15
EuroSAT	75.81	82.67	86.89	98.00	98.44	97.63	97.07	97.37	98.37	98.26	97.56	98.41	98.59	99.89
GTSRB	72.01	61.93	70.86	94.48	95.36	94.09	95.71	94.86	96.63	97.05	91.82	95.32	97.11	99.24
MNIST	98.05	93.83	98.33	99.26	99.24	99.46	99.57	99.54	99.59	99.65	98.97	99.29	99.64	99.69
RESISC45	76.08	78.71	82.70	94.22	94.78	92.49	92.06	94.10	93.10	93.30	92.67	95.25	94.81	97.37
SVHN	81.91	72.97	84.18	91.07	90.16	93.79	93.73	95.24	94.98	95.81	86.14	90.48	93.53	98.11
SUN397	69.01	70.95	72.92	78.62	78.94	76.17	76.96	78.92	77.77	78.33	75.58	79.46	78.21	82.32
STL10	97.50	99.38	98.51	99.54	99.61	99.44	99.46	99.42	99.52	99.50	99.65	99.60	99.59	99.38
OxfordIIITPet	93.98	94.74	94.58	96.48	96.40	96.21	96.21	96.08	96.32	96.32	96.48	96.48	95.99	95.15
Flowers102	69.85	80.55	65.52	89.92	96.86	87.64	90.18	78.81	93.23	88.06	89.51	96.62	96.06	90.88
CIFAR100	74.46	82.91	80.21	87.87	88.35	85.40	85.54	86.82	86.90	86.31	86.37	89.01	88.28	91.97
PCAM	79.58	74.38	80.35	82.08	81.08	84.06	84.10	83.34	84.75	85.40	86.41	80.01	83.68	90.07
FER2013	46.99	42.45	50.06	57.11	58.48	67.89	67.39	73.18	71.32	73.57	64.78	60.27	74.58	75.16
Average	75.62	76.76	78.65	88.02	88.95	88.24	88.64	88.80	89.91	89.87	87.59	89.19	90.73	92.55

Table 15: **Per-task test accuracy (%) for ViT-L/14 with 20 tasks.** The final row reports the unweighted mean over the 20 tasks.

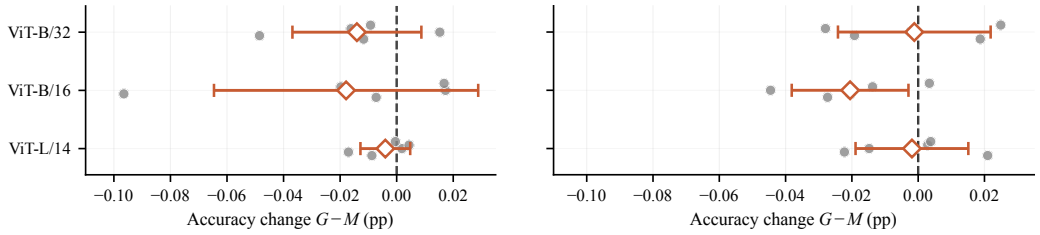
Dataset	TA	Avg	TIES	ISO-C	ISO-CTS	TSV-M	WUDI	ACTMat	SWUDI-A	ACE	DC	SVC	SOS-MERGE	Fine-tuned
Cars	40.08	78.80	66.96	76.98	85.75	80.81	76.00	84.99	85.57	85.51	86.93	85.09	87.91	92.39
DTD	42.34	58.56	56.01	75.16	77.93	72.34	65.80	75.16	74.89	74.84	70.69	78.35	79.10	84.15
EuroSAT	47.56	77.11	71.44	96.56	98.44	95.67	81.67	96.19	97.74	97.56	96.93	98.41	98.22	99.89
GTSRB	46.68	59.22	62.98	91.88	94.77	90.70	85.51	90.51	94.17	93.40	89.38	94.20	94.88	99.24
MNIST	94.91	92.63	96.43	98.21	98.57	97.25	96.95	95.05	97.67	96.66	96.81	98.48	98.09	99.69
RESISC45	44.73	76.46	74.29	92.60	94.05	90.22	84.03	92.43	91.89	92.19	91.76	94.27	94.14	97.37
SVHN	67.80	70.89	79.82	88.31	88.42	90.08	85.69	91.99	92.07	92.60	82.47	87.70	90.09	98.11
SUN397	56.14	70.43	70.91	77.22	78.21	74.76	73.10	77.37	76.46	76.75	74.93	78.69	77.16	82.32
STL10	90.76	99.38	97.69	99.30	99.59	99.22	98.45	99.29	99.48	99.42	99.60	99.56	99.60	99.38
OxfordIIITPet	85.80	94.28	93.79	95.78	96.02	95.94	95.09	95.97	96.27	96.05	96.46	96.18	96.32	95.15
Flowers102	45.83	79.93	64.25	85.44	96.34	85.12	80.74	78.08	91.41	85.33	88.97	96.03	95.22	90.88
CIFAR100	56.04	82.62	75.20	84.90	87.05	82.45	76.58	84.72	85.15	83.74	84.95	87.72	86.78	91.97
PCAM	68.41	71.06	75.23	83.03	81.97	85.33	81.68	84.42	84.96	86.02	86.40	81.61	84.89	90.07
FER2013	40.32	40.54	44.96	54.82	57.02	62.62	54.76	69.60	65.31	69.03	60.77	57.29	71.23	75.16
CIFAR10	88.61	97.27	96.06	98.13	98.69	97.98	96.53	98.14	98.39	98.14	98.22	98.82	98.62	99.13
Food101	47.35	90.66	82.70	90.44	92.18	90.53	86.01	91.78	92.03	92.00	92.69	92.24	92.99	92.17
FashionMNIST	75.72	76.24	82.59	90.43	90.95	90.69	88.98	90.23	90.53	91.02	89.18	90.99	91.22	95.67
RenderedSST2	62.27	67.38	68.09	80.40	82.48	83.53	81.55	85.50	85.23	85.78	84.07	83.20	85.23	84.40
EMNIST	70.32	36.55	88.75	96.88	94.69	99.23	98.75	99.59	99.07	99.54	96.05	93.99	98.91	99.84
KMNIST	26.54	11.42	34.84	80.37	80.22	79.07	62.07	78.69	75.52	80.05	69.60	77.99	83.95	98.72
Average	59.91	71.57	74.15	86.84	88.67	87.18	82.50	87.98	88.69	88.78	86.84	88.54	90.23	93.28

where $M_{i\ell}$ captures the magnitude of the retained spectrum and $H_{i\ell}$ captures its normalized shape. To isolate their effects, all variants use the same support $P_{i\ell}$, and their layerwise total scalar weight is matched to that of the full $G_{i\ell}$ configuration.

Tab. 16 first evaluates this decomposition on the canonical 8-, 14-, and 20-task settings. The full scalar $G = MH$ achieves the highest average accuracy of 86.92%, compared with 86.86% for M

Table 16: **Decomposition of the scalar spectral weight under matched layerwise total scale.** All variants use the same support $P_{i\ell}$. The full scalar $G = MH$ achieves the highest average accuracy, while the magnitude-only component M remains close.

	ViT-B/32			ViT-B/16			ViT-L/14			Average
Scale on $P_{i\ell}$	8	14	20	8	14	20	8	14	20	↑
$M_{i\ell}$	**86.40**	83.01	80.13	**89.40**	86.64	83.08	92.44	90.66	89.98	86.86
$H_{i\ell}$	86.28	82.77	79.97	89.37	86.45	83.10	92.40	90.68	90.12	86.79
$G_{i\ell} = M_{i\ell}H_{i\ell}$	86.39	**83.05**	**80.18**	89.40	**86.65**	**83.17**	**92.45**	**90.73**	**90.23**	**86.92**



(a) $K = 8$: $5/15$ positive.

(b) $K = 14$: $6/15$ positive.

Figure 3: **Effect of the normalized-shape component across random fixed-size task subsets.** For each $K \in \{8, 14\}$, five fixed subsets are sampled from the 20-task pool and reused across all three backbones. Each gray point shows the paired macro-accuracy difference $G - M$ for one subset–backbone pair. The vertical dashed line marks zero, and diamonds with horizontal whiskers show the backbone average and sample standard deviation.

and 86.79% for H. The advantage of G over M is therefore only 0.05 percentage points on average. These results suggest that most of the benefit of the scalar weight is already captured by spectral magnitude, while the normalized-shape component provides at most a small additional refinement on the canonical task prefixes.

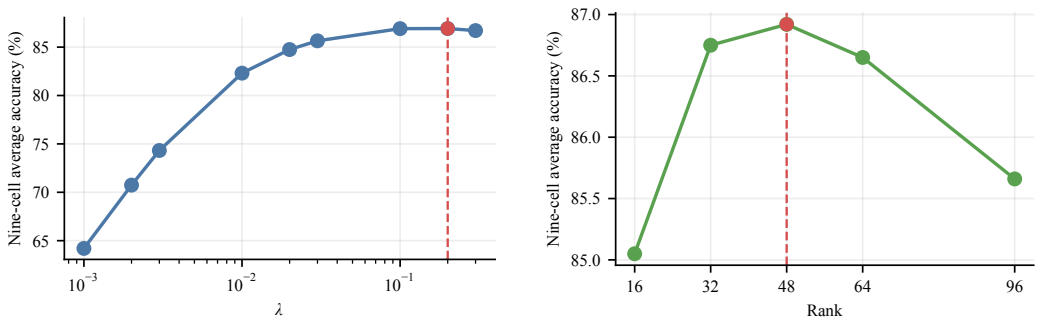
We next test whether this small difference is stable under changes in task composition. For each $K \in \{8, 14\}$, we sample five task subsets without replacement from the same 20-task pool, fix these subsets before evaluation, and reuse the identical subset manifest for ViT-B/32, ViT-B/16, and ViT-L/14. We compare the full scalar G with the magnitude-only scalar M using the paired macro-accuracy difference $G - M$ for each subset–backbone pair.

As shown in Fig. 3, the difference is close to zero and slightly negative when pooled across the 15 subset–backbone pairs. For $K = 8$, the average difference is -0.012 points with 5 of 15 pairs positive; for $K = 14$, it is -0.008 points with 6 of 15 pairs positive. Thus, the small benefit of the normalized-shape component observed on the canonical prefixes does not persist consistently across task compositions. The magnitude component M accounts for the more stable part of the scalar spectral weight.

C.3 Hyperparameter Sensitivity

We examine the sensitivity of SOS-Merge to the two main numerical hyperparameters, the ridge coefficient λ and the retained rank r. These sweeps use test accuracy only for post-hoc diagnosis and are not used to select the configuration reported in the main experiments. Each point averages over the same nine backbone–task-count settings as Tab. 4.

As shown in Fig. 4, very small ridge values substantially reduce performance, while $\lambda = 0.1$ and the default $\lambda = 0.2$ achieve similar average accuracies of 86.91% and 86.92%, respectively. Performance is also relatively stable for retained ranks between 32 and 64, with rank 48 giving the highest nine-setting average. These results indicate that the default configuration is not sensitive to small changes around either hyperparameter, while insufficient regularization is the main source of degradation in the ridge sweep.



(a) Ridge coefficient λ.

(b) Retained rank.

Figure 4: **Sensitivity to the ridge coefficient and retained rank.** Each point reports the average test accuracy over the nine backbone–task-count settings. The default configuration uses $\lambda = 0.2$ and rank 48. These sweeps are diagnostic and are not used for model selection.

Table 17: **Comparison with baselines using validation-based coefficient selection.** Entries are macro test top-1 accuracies (%), and parentheses denote the selected coefficient. ACE-Merging and SOS-MERGE retain their data-free configurations and therefore do not use validation-based coefficient selection.

	ViT-B/32			ViT-B/16			ViT-L/14			Average
Method	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	↑
Fine-tuned	90.53	89.20	89.99	92.61	91.21	91.94	94.14	92.55	93.28	91.72
Task Arithmetic	68.95 (0.3)	65.32 (0.1)	60.40 (0.1)	74.02 (0.3)	70.36 (0.1)	65.56 (0.1)	84.35 (0.3)	79.35 (0.2)	73.99 (0.1)	71.37
TIES global	73.53 (0.9)	66.93 (0.8)	61.05 (0.8)	78.05 (0.9)	72.27 (0.8)	67.26 (0.8)	85.88 (1.0)	79.14 (0.9)	75.46 (0.8)	73.29
ISO-C	83.88 (1.4)	78.74 (1.0)	74.53 (0.8)	88.20 (1.4)	83.37 (1.0)	78.28 (0.7)	91.89 (2.1)	88.14 (1.1)	86.40 (1.2)	83.71
ISO-CTS	84.06 (1.5)	79.90 (1.4)	76.79 (1.1)	89.05 (1.5)	84.70 (1.3)	81.49 (1.0)	92.51 (2.4)	89.51 (1.6)	89.15 (1.4)	85.24
TSV-M	83.54 (1.1)	79.37 (1.0)	76.34 (0.8)	86.71 (1.2)	83.83 (1.0)	79.92 (1.0)	91.23 (1.3)	88.24 (1.0)	87.18 (1.0)	84.04
DC-Merge	85.08 (1.7)	81.66 (1.6)	79.68 (1.9)	88.62 (2.0)	85.91 (1.7)	**83.34** (2.1)	92.36 (2.3)	89.76 (2.4)	89.87 (1.9)	86.25
ACE-Merging	85.41	81.24	78.04	89.10	85.20	81.52	**92.54**	89.87	88.78	85.74
SOS-MERGE	**86.39**	**83.05**	**80.18**	**89.40**	**86.65**	83.17	92.45	**90.73**	**90.23**	**86.92**

C.4 FEW-SHOT COMPARISONS

Our main evaluation uses strictly data-free configurations. As an additional reference, we consider baselines that support validation-based coefficient selection and report their performance after such selection. The selected coefficients are shown in parentheses in Tab. 17. ACE-Merging and SOS-MERGE retain the same data-free configurations as in the main experiments, providing direct reference points without additional model selection. ACTMat and SWUDI-A are not repeated because their configurations are unchanged from the main data-free comparison.

Allowing validation-based coefficient selection improves several baselines relative to their fixed data-free configurations, particularly DC-Merge. Nevertheless, SOS-MERGE retains the highest nine-setting average at 86.92% without using validation data for coefficient selection. DC-Merge reaches 86.25% after validation-based tuning and exceeds SOS-MERGE in one setting, ViT-B/16 with 20 tasks. These results provide an additional reference for the performance gap between the strictly data-free setting and baselines that can adapt their merging coefficient using validation information.