RELEVANCE MEETS SPECTRUM: JOINT CALIBRATION FOR MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Singular value decomposition (SVD) has recently emerged as a key technique in several strong methods for model merging. In particular, existing SVD-based methods organize task-related directional structures and spectra in different ways, with the general goal of preserving useful information from multiple task-specific updates. However, a single merged update generally cannot preserve the full directional structures of all task-specific updates simultaneously. Motivated by this issue, we prioritize task-relevant structure within the dominant low-rank representation of the merged update, rather than seeking complete directional alignment. This prioritization also makes spectral allocation important, since the singular values determine how strongly the retained directions contribute to the merged update. Based on these considerations, we propose RSS-Merge, a model-merging method that combines relevance-oriented directional construction with order-preserving spectrum calibration. Extensive experiments show that RSS-Merge achieves the best average accuracy across all nine evaluated vision settings. On the vision-language MM-MergeBench, it obtains the highest unseen-task average and ranks second on seen tasks, only marginally behind the best result.

1 INTRODUCTION

Large pretrained models have become a standard starting point for modern learning systems (Radford et al., 2021; Zhai et al., 2023). Fine-tuning such a model for an individual downstream task often produces a specialized model with strong performance (Wortsman et al., 2022b; Ilharco et al., 2022), and the growing range of downstream applications has led to a large collection of publicly available task-specific models. However, when multiple capabilities are required in one system, maintaining these models separately incurs storage and deployment costs that grow with the number of tasks. Moreover, training a unified multi-task model requires additional optimization and typically assumes access to the original data. Model merging offers a different approach: it directly consolidates multiple task-specific models into a single model without revisiting their training data or performing additional training (Li et al., 2023; Wortsman et al., 2022a; Ilharco et al., 2023). This combination of efficiency and flexibility has motivated growing interest in model merging (Yadav et al., 2023; Yang et al., 2024; Gargiulo et al., 2025).

Among recent advances, singular value decomposition (SVD) has emerged as a powerful tool for exploiting the matrix structure of task-specific updates. In particular, SVD decomposes each task update into singular vectors that characterize the directional geometry of task information and singular values that determine the relative strength of singular vectors. Existing methods manipulate these components in different ways, including retaining dominant task-specific components (Choi et al., 2024; Gargiulo et al., 2025), reducing overlap among task subspaces (Gargiulo et al., 2025), aligning low-rank updates in joint or shared cover spaces (Stoica et al., 2025; Zhang et al., 2026), combining common and task-specific structures (Marczak et al., 2025), and reshaping singular values before constructing the merged update (Zeng et al., 2025; Marczak et al., 2025; Zhang et al., 2026). Despite their different implementations, these methods share a natural objective: the merged representation should preserve useful directional information from the constituent tasks.

Note that task-specific updates can exhibit distinct and potentially incompatible directional geometries, so a single merged update cannot preserve the full structures of all constituent updates simultaneously. Motivated by this incompatibility, we focus on making task-relevant structure prominent

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.

in the dominant low-rank representation of the merged update, namely its leading singular subspace, rather than seeking complete directional alignment across the entire update. From the perspective of SVD, however, directional geometry is only one factor governing how information is expressed. Another factor is how spectral strength is distributed across the constructed directions. Singular values assign relative strength to the singular vectors and thereby determine this distribution. A highly non-uniform spectrum may place disproportionate emphasis on a few directions, diminishing the contribution of other task-relevant or complementary directions.

Based on these considerations, we propose Relevance Meets Spectrum: Joint Calibration for Model Merging (RSS-Merge), a model-merging method that jointly shapes the directional and spectral structure of the merged update. Specifically, RSS-Merge first centers each task update by the average update and aggregates its leading low-rank residual components. This residual branch exposes salient task-relative variations and anchors the leading singular subspace. The resulting relevance-oriented update is then supplemented with the component of the average update that is orthogonal to the residual subspaces on both the left and right, recovering complementary information while leaving the residual directions unchanged. RSS-Merge subsequently calibrates the spectrum by smoothing the singular values. Since their ordering determines which directions occupy the leading singular subspace, the calibration preserves this ordering and thereby retains the relevance-oriented geometry. Across nine vision settings covering three backbones and three task-set sizes, RSS-Merge achieves the best average accuracy in every setting. On MM-MergeBench, it obtains the highest unseen-task average of 40.07% and the second-highest seen-task average of 59.20%, only 0.43 percentage points below the best result.

The main contributions of this paper are summarized as follows:

- We formulate a leading-subspace principle for model merging: when a merged update cannot preserve the full directional geometry of multiple distinct task updates simultaneously, task relevance should be concentrated in its leading singular subspace.
- We propose RSS-Merge, which constructs a relevance-oriented leading geometry from salient task-relative variations and complementary information from the average update, then calibrates its spectrum without changing the constructed directions or their ordering.
- We evaluate RSS-Merge on vision and vision-language benchmarks. It achieves the best average accuracy in all nine vision settings, as well as the best unseen-task average and second-best seen-task average on MM-MergeBench; controlled ablations verify the complementary roles of its two components.

2 Related Work

Early work on model merging focused primarily on combining models directly in parameter space. Model Soups (Wortsman et al., 2022a) showed that averaging independently fine-tuned models can improve accuracy and robustness without increasing inference cost, while Fisher Merging (Matena & Raffel, 2022) accounted for differences in parameter importance through Fisher-weighted averaging. Task Arithmetic (TA) (Ilharco et al., 2023) subsequently represented each fine-tuning update as a task vector relative to the pretrained model, enabling arithmetic composition of task-specific capabilities. Subsequent methods build on this task-vector formulation to improve how task updates are selected, weighted, or combined. TIES-Merging (Yadav et al., 2023) resolves sign conflicts after removing low-magnitude updates, DARE (Yu et al., 2024) sparsifies and rescales task vectors, and Consensus Merging (Wang et al., 2024), PCB-Merging (Du et al., 2024), and WUDI-Merging (Cheng et al., 2025) localize task information or suppress competing updates. These methods improve on plain task-vector aggregation by mitigating interference at the parameter-coordinate level, but they use no task data or test-time adaptation to determine the merge. A separate line of work uses additional statistics or data to adapt the merge: RegMean (Jin et al., 2023) uses activation statistics to solve for layer-wise merged weights, while AdaMerging (Yang et al., 2024) learns task-wise or layer-wise mixing coefficients from unlabeled test data. Collectively, these parameter-space approaches refine which coordinates are retained and how they are weighted, but they do not explicitly engage with the matrix-level structure of task updates: they neither use SVD to characterize task-update subspaces nor control how information is distributed across those subspaces and their spectra.

More recently, a growing line of work has leveraged SVD to explicitly model the directional and spectral structure of task updates. KnOTS (Stoica et al., 2025) aligns low-rank task updates in a joint space before merging. CART (Choi et al., 2024) revisits task arithmetic from the perspective of weight averaging: rather than decomposing updates relative to the pretrained model, it takes the average update as the reference, applies truncated SVD to the resulting centered task residuals, and adds the retained components back to the average update. This re-centering reduces interference among task updates, while rank reduction retains the dominant components that carry most task-specific knowledge, yielding improved performance over weight averaging and standard task-arithmetic baselines. TSV-Merge (Gargiulo et al., 2025) further retains dominant task components while orthogonalizing their singular vectors to reduce subspace overlap. RobustMerge (Zeng et al., 2025) corrects singular-value disparities in parameter-efficient updates, while DC-Merge (Zhang et al., 2026) smooths task-update spectra and aligns their directions in a shared orthogonal cover space. Iso-CTS (Marczak et al., 2025) is the closest comparison to RSS-Merge because both methods jointly shape the directional composition and spectral allocation of the merged update. They differ, however, in both objective and construction. Iso-CTS seeks a balanced representation of tasks: it flattens the spectrum of the common subspace to reduce its bias toward well-aligned tasks and introduces task-specific directions for tasks that would otherwise be underrepresented. RSS-Merge instead seeks to concentrate task-relevant information in the dominant low-rank representation. It constructs this representation from centered task residuals and complementary average-update structure, then moderates spectral concentration so that its information is not dominated by only a few directions.

3 METHOD: RSS-MERGE

In this section, we present RSS-Merge, which jointly shapes the directional relevance and spectral strength of the merged update. At a high level, it can be understood as a two-stage framework: relevance construction establishes the directional structure, and order-preserving spectrum calibration regulates the associated spectral strength.

3.1 RELEVANCE CONSTRUCTION

Task-relative Directional Anchor. We aim to construct a merged update whose leading singular subspace is representative of all constituent tasks. TA (Ilharco et al., 2023) is a natural starting point because its aggregate update combines information from all tasks and its leading singular subspace captures a shared structure across their updates. However, this aggregate construction does not guarantee balanced task coverage. Iso-CTS (Marczak et al., 2025) shows that the leading subspace of the TA update preferentially represents tasks whose task matrices are well aligned with the common subspace: their updates contribute more coherently to the dominant directions, whereas less-aligned or more task-specific directions can remain present in the full aggregate update but be underrepresented in its leading subspace. Thus, an update can contain information from every task while its dominant low-rank representation remains biased toward the common structure.

To make the leading subspace more representative, we need a construction that reduces the influence of common-subspace alignment and gives salient task-relative variations sufficient representation. CART (Choi et al., 2024) offers a concrete mechanism for this purpose: it expresses each task update relative to the average update and applies a truncated SVD to the resulting residual. Centering exposes task-relative deviations, while truncation retains their principal task-specific structure. We therefore use centered residual truncation as the residual branch of RSS-Merge, directing the leading subspace toward task-relevant variations rather than allowing common-subspace alignment in the aggregate update to determine which directions are promoted. Let ΔW_t denote the update for task t relative to the pretrained model, and let $\overline{\Delta W}$ denote their average. We retain the leading p fraction of singular components from each centered task update and aggregate them as

$$\Delta_{\text{res}} = \sum_{t=1}^{T} \text{SVD}_p\big(\Delta W_t - \overline{\Delta W}\big), \qquad \overline{\Delta W} = \frac{1}{T}\sum_{t=1}^{T} \Delta W_t. \tag{1}$$

Here, T is the number of task-specific models, and SVD_p reconstructs each centered update from its leading p fraction of singular components, with at least one component retained. This residual

branch provides the task-relative directional anchor; the complementary average-update structure introduced next restores information outside the selected residual subspaces.

Complementary average-update structure. The residual branch is intentionally low-rank: it prioritizes a small set of task-relative directions and does not capture all information in the task updates. The remaining space should therefore retain useful shared or complementary structure from the average update. A straightforward approach would be to add the average update directly. In general, however, the average update has nonzero projections onto the residual left and right singular subspaces. Direct addition would therefore couple the two branches in the SVD, potentially rotating the singular directions that define the task-relative anchor. We instead seek an average-update component that is orthogonal to the residual subspaces on both the left and right. This bilateral constraint keeps the two branches in disjoint row and column subspaces, allowing the residual directions to remain unchanged while retaining the part of the average update that is compatible with them. Let U_{res}, Σ_{res}, and V_{res} denote the compact SVD factors of the residual update, restricted to its nonzero spectrum. We obtain the complementary component by bilateral projection:

$$\Delta^{\perp}_{\text{mean}} = (I - U_{\text{res}}U_{\text{res}}^{\top})\overline{\Delta W}(I - V_{\text{res}}V_{\text{res}}^{\top}). \tag{2}$$

By construction, the projected average update is orthogonal to the residual singular subspaces on both the left and right. Among all matrices satisfying this bilateral orthogonality requirement, it is the closest approximation to the original average update in Frobenius norm. It therefore retains as much of the average update as the constraints allow. We combine the two branches as

$$\Delta_{\text{rel}} = \Delta^{\perp}_{\text{mean}} + \lambda\Delta_{\text{res}}, \tag{3}$$

where $\lambda > 0$ controls the relative spectral strength of the residual branch. To see how the two branches interact, let $U_{\perp}$, $\Sigma_{\perp}$, and $V_{\perp}$ denote the compact SVD factors of the complementary component. The bilateral projection makes its left and right singular subspaces orthogonal to their residual counterparts. Therefore we have

$$\Delta_{\text{rel}} = \begin{bmatrix} U_{\text{res}} & U_{\perp} \end{bmatrix} \begin{bmatrix} \lambda\Sigma_{\text{res}} & 0 \\ 0 & \Sigma_{\perp} \end{bmatrix} \begin{bmatrix} V_{\text{res}} & V_{\perp} \end{bmatrix}^{\top}, \tag{4}$$

which is an SVD up to component reordering. Consequently, combining the two branches preserves their singular directions exactly. The coefficient λ rescales only the residual spectrum, while the joint spectrum of the two branches determines which directions enter a fixed leading band.

3.2 Order-Preserving Spectrum Calibration

Relevance construction determines which directions should define the merged update, but it does not by itself determine how strongly those directions contribute. The orthogonal residual and complementary branches can still induce a concentrated spectrum in which a few components dominate the update magnitude and other task-relevant directions receive comparatively little strength. Directional relevance alone is therefore insufficient: the selected geometry must also be accompanied by a less concentrated allocation of spectral strength. We consequently calibrate the nonzero singular values while preserving their ordering, so that the directional priority established above is retained and only the relative strength of the directions is moderated.

Let U_{rel}, V_{rel}, and s_{rel} denote the compact SVD factors of Δ_{rel}, restricted to its nonzero spectrum, and let $\bar{s}$ denote the mean of the entries in s_{rel}. We move each singular value toward the mean using

$$s^{(\gamma)}_{\text{rel},i} = (1-\gamma)s_{\text{rel},i} + \gamma\bar{s}, \qquad 0 \le \gamma < 1. \tag{5}$$

This affine transformation has the mean as its fixed point and a positive slope smaller than one. It therefore contracts deviations from the mean without reversing the ordering of the nonzero singular values. Because the singular vectors are kept fixed, the directions and their selection by any fixed leading-rank cutoff are unchanged, up to pre-existing ties.

Substituting the calibrated spectrum from equation 5 into the compact SVD of Δ_{rel} yields the following matrix-level relation:

$$\begin{aligned} \Delta &= U_{\text{rel}}\,\text{diag}(s^{(\gamma)}_{\text{rel}})V_{\text{rel}}^{\top} \\ &= (1-\gamma)\Delta_{\text{rel}} + \gamma\bar{s}U_{\text{rel}}V_{\text{rel}}^{\top}. \end{aligned} \tag{6}$$

Algorithm 1 RSS-Merge

Input: Task vectors $\{\Delta W_t\}_{t=1}^{T}$ with $\Delta W_t \in \mathbb{R}^{m\times n}$; pretrained matrix W_0; residual weight λ; retention ratio p; smoothing coefficient γ; update scale α
`// Stage 1: Relevance construction`
Compute the average update $\overline{\Delta W} \leftarrow T^{-1}\sum_{t=1}^{T}\Delta W_t$
Set truncation rank $k \leftarrow \max\{1, \lfloor p\min(m,n)\rfloor\}$
Initialize $\Delta_{\mathrm{res}} \leftarrow 0_{m\times n}$
for $t = 1, \ldots, T$ **do**
    Center: $\delta_t \leftarrow \Delta W_t - \overline{\Delta W}$
    Decompose: $(U_t, s_t, V_t) \leftarrow \mathrm{SVD}(\delta_t)$
    Accumulate rank-k update: $\Delta_{\mathrm{res}} \leftarrow \Delta_{\mathrm{res}} + U_{t,:k}\,\mathrm{diag}(s_{t,:k})V_{t,:k}^{\top}$
end for
Nonzero compact SVD: $(U_{\mathrm{res}}, s_{\mathrm{res}}, V_{\mathrm{res}}) \leftarrow \mathrm{SVD}(\Delta_{\mathrm{res}})$
Bilateral projection: $\Delta_{\mathrm{mean}}^{\perp} \leftarrow (I - U_{\mathrm{res}}U_{\mathrm{res}}^{\top})\overline{\Delta W}(I - V_{\mathrm{res}}V_{\mathrm{res}}^{\top})$
Combine: $\Delta_{\mathrm{rel}} \leftarrow \Delta_{\mathrm{mean}}^{\perp} + \lambda\Delta_{\mathrm{res}}$
`// Stage 2: Spectrum calibration`
Nonzero compact SVD: $(U_{\mathrm{rel}}, s_{\mathrm{rel}}, V_{\mathrm{rel}}) \leftarrow \mathrm{SVD}(\Delta_{\mathrm{rel}})$
Compute mean singular value $\bar{s} \leftarrow \mathrm{mean}(s_{\mathrm{rel}})$
Order-preserving smoothing: $s_{\mathrm{rel}}^{(\gamma)} \leftarrow (1-\gamma)s_{\mathrm{rel}} + \gamma\bar{s}\mathbf{1}_{r_{\mathrm{rel}}}$
Reconstruct: $\Delta \leftarrow U_{\mathrm{rel}}\,\mathrm{diag}(s_{\mathrm{rel}}^{(\gamma)})V_{\mathrm{rel}}^{\top}$
$W_{\mathrm{merge}} \leftarrow W_0 + \alpha\Delta$
return W_{merge}

Thus, at the matrix level, calibration interpolates between the relevance-constructed update and its equal-spectrum counterpart while retaining the same left and right singular subspaces. The merged model is obtained by adding the calibrated update to the pretrained weights, where α is a global rescaling coefficient:

$$W_{\mathrm{merge}} = W_0 + \alpha\Delta. \tag{7}$$

4 Experiments

In this section, we first evaluate the performance of RSS-Merge against existing model-merging baselines on vision and vision-language benchmarks. We then conduct ablation studies to examine the contributions of relevance construction and spectrum calibration, as well as the sensitivity of RSS-Merge to its key design choices.

4.1 Vision Model Merging

In this subsection, we investigate fully fine-tuned vision model merging under the benchmark setting established by Wang et al. (2024). Specifically, we evaluate the 8-task, 14-task, and 20-task configurations, which provide progressively larger task collections for assessing the behavior of model merging as the number of merged tasks increases. The vision models use CLIP (Radford et al., 2021) with ViT-B/32, ViT-B/16, and ViT-L/14 as the visual encoders (Dosovitskiy et al., 2021). All implementation details are provided in Appendix A.

We include the zero-shot and individual task-specific models as reference points, and compare against TA (Ilharco et al., 2023), DC-Merge (Zhang et al., 2026), TIES-Merging (Yadav et al., 2023), Consensus TA (Wang et al., 2024), TSV-M (Gargiulo et al., 2025), and Iso-CTS (Marczak et al., 2025). Following prior work (Wang et al., 2024; Gargiulo et al., 2025; Marczak et al., 2025), we report both average absolute accuracy and average normalized accuracy. RSS-Merge achieves the best performance across all evaluated backbone and task-count settings among the merging methods, and its advantage is maintained as the number of merged tasks increases from 8 to 20.

4.2 Vision-Language Model Merging

We next study vision-language model merging on MM-MergeBench. The benchmark contains eight seen tasks used to construct the merged model and four additional unseen tasks used to evaluate transfer beyond the task set used during merging. Following the original experimen-

Table 1: Average absolute accuracy (%) on the standard vision model merging benchmark in the FFT setting. The value in parentheses is the average normalized accuracy (%). The best results are in bold and the second-best are underlined. We evaluate three CLIP backbones and 8-task, 14-task, and 20-task configurations.

Method	ViT-B/32			ViT-B/16			ViT-L/14		
	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
Zero-shot	48.41	57.28	56.15	55.38	61.28	59.71	65.00	68.37	65.35
Individual	93.31	91.16	91.57	94.95	92.94	93.30	96.12	94.48	94.86
TA	69.12 (73.77)	65.24 (71.71)	60.47 (66.53)	75.43 (78.97)	70.52 (75.66)	65.78 (70.53)	84.93 (88.35)	79.41 (82.58)	74.01 (77.91)
TIES-Merging	73.40 (79.23)	67.69 (74.49)	64.84 (71.35)	78.16 (82.68)	72.85 (78.32)	70.11 (75.21)	85.08 (88.84)	80.33 (84.94)	77.13 (81.25)
ConsensusTA	74.40 (80.13)	70.08 (76.99)	67.31 (73.88)	79.35 (83.75)	74.16 (79.65)	72.24 (77.41)	86.23 (89.98)	82.33 (87.03)	80.06 (84.32)
TSV-M	85.48 (91.95)	80.02 (87.86)	76.82 (84.06)	88.75 (93.70)	84.43 (90.88)	80.63 (86.52)	92.93 (96.96)	89.16 (94.45)	87.57 (92.36)
Iso-CTS	86.45 (93.15)	81.46 (89.44)	78.31 (85.71)	91.18 (96.28)	86.27 (92.76)	82.84 (88.83)	94.71 (98.83)	91.17 (96.50)	90.24 (95.09)
DC-Merge	87.19 (93.24)	82.51 (90.29)	80.60 (88.15)	90.57 (95.25)	86.78 (93.26)	84.31 (90.23)	94.20 (98.29)	90.88 (96.32)	90.27 (95.27)
RSS-Merge	**88.76 (95.05)**	**84.14 (92.20)**	**81.23 (88.73)**	**91.96 (96.80)**	**87.31 (93.80)**	**84.43 (90.48)**	**94.73 (98.95)**	**91.68 (96.91)**	**90.98 (95.83)**

Table 2: Task-wise accuracy (%) on MM-MergeBench (Zeng et al., 2025) using LLaVA-v1.5-7B with LoRA task vectors. The first group contains eight seen tasks and the second contains four unseen tasks. Individual, zero-shot, and multi-task results are reference points; among merging methods, the best and second-best results are shown in bold and underlined, respectively.

Method	Seen Tasks									Unseen Tasks				
	SciQA	Image	VQA	REC	OCR	VizWiz	Flickr	IconQA	Avg.	AVQA	Image-R	S2W	TAbMWP	Avg.
Individual	83.74	96.02	67.58	43.40	65.50	64.80	57.29	75.54	69.23	–	–	–	–	–
Zero-shot	61.73	40.87	62.88	36.10	41.16	41.03	49.07	14.09	43.37	51.62	28.27	5.98	15.01	25.22
Multi-task	76.90	74.08	67.05	35.98	65.37	66.67	56.09	66.87	63.63	76.33	41.39	8.34	18.20	36.06
TA	71.94	57.49	67.06	38.90	62.87	44.80	49.20	39.21	53.93	74.78	37.37	7.52	13.57	33.31
DARE-Merging	71.59	57.25	66.26	39.38	62.56	44.93	49.13	39.59	53.84	73.75	37.67	7.56	13.62	33.15
TIES-Merging	71.49	55.88	66.73	39.67	**65.12**	44.35	47.06	34.46	53.10	73.43	38.44	7.47	13.23	33.14
PCB-Merging	71.10	57.82	**67.59**	38.22	64.35	44.58	48.90	37.01	53.70	74.57	36.28	7.84	15.44	33.53
TSV-M	75.66	**78.61**	59.48	41.92	41.71	41.19	52.23	45.97	54.60	77.29	42.72	11.12	13.80	36.23
WUDI-Merging	69.46	78.40	56.84	38.95	38.58	40.23	51.13	38.82	51.55	68.56	41.57	11.16	11.80	33.27
Iso-CTS	77.54	77.01	63.50	45.72	40.75	42.34	54.74	53.25	56.86	78.48	44.81	**11.21**	13.99	37.12
RobustMerge	73.43	65.54	67.20	44.80	62.97	**46.61**	52.80	45.90	57.41	79.30	45.79	9.23	17.62	37.99
DC-Merge	77.05	73.33	65.87	**47.88**	59.66	45.23	**55.06**	52.92	**59.63**	79.30	50.24	10.88	**18.94**	39.84
RSS-Merge	**78.33**	73.49	66.24	45.47	56.43	45.64	54.50	**53.53**	59.20	**79.83**	**50.89**	10.66	18.88	**40.07**

tal setup (Zeng et al., 2025), we use the LLaVA-v1.5-7B backbone (Liu et al., 2023) and merge LoRA (Hu et al., 2022) task-specific checkpoints.

To ensure a matched comparison, we use the LLaVA-v1.5-7B checkpoints released by Robust-Merge (Zeng et al., 2025). We compare RSS-Merge against TA (Ilharco et al., 2023), DARE-Merging (Yu et al., 2024), TIES-Merging (Yadav et al., 2023), PCB-Merging (Du et al., 2024), TSV-M (Gargiulo et al., 2025), WUDI-Merging (Cheng et al., 2025), Iso-CTS (Marczak et al., 2025), RobustMerge (Zeng et al., 2025), and DC-Merge (Zhang et al., 2026) under the same task partition.

We report task-wise performance together with the average performance of the seen-task and unseen-task groups. Table 2 reports the detailed results. Among the merging methods, RSS-Merge achieves the highest scores on SciQA and IconQA among the seen tasks and on two of the four unseen-task metrics (AVQA and Image-R). Its group averages are 59.20% on the seen tasks and 40.07% on the unseen tasks.

4.3 Analysis and Ablations

We first analyze the directional structure of the final update under a fixed leading-band capacity. We then isolate the effects of relevance construction and spectrum calibration, followed by an examination of the complementary average-update branch and its bilateral projection. Finally, we study the effect of the residual retention ratio p to justify the choice $p = 1\%$ within the tested range, and assess the robustness of RSS-Merge to the residual weight λ. Unless stated otherwise, the analyses in this subsection use fully fine-tuned vision models with the same task checkpoints, evaluation protocol, and held-out validation split as in Section 4.1. For every variant and hyperparameter setting, the global rescaling coefficient α is selected independently using the same validation protocol.

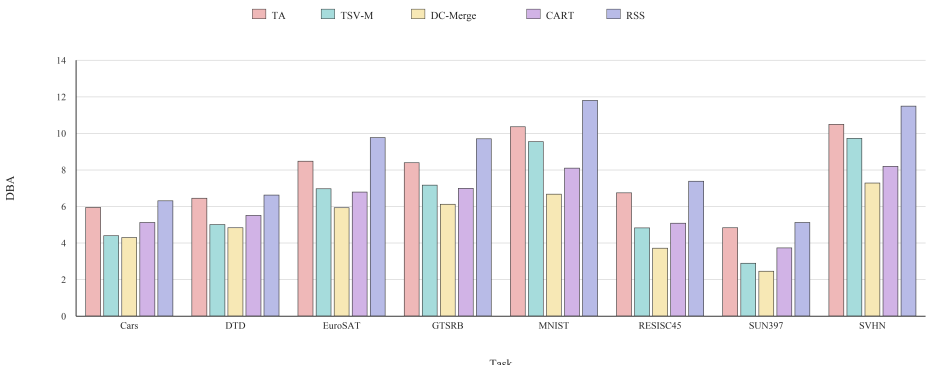


Figure 1: DBA comparison across 8 tasks on ViT-B/32.

4.3.1 Leading-Subspace Directional Analysis

We use a structural diagnostic to test whether task-relevant information is concentrated among the directions that dominate the merged update. A fair comparison requires the same directional capacity for every method: a wider subspace can capture more task-update energy simply by containing more directions. Let $M \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be a matrix-valued merged update with SVD $M = U \operatorname{diag}(s) V^\top$, where the singular values are in non-increasing order. For a relative band ratio ρ, we set $k = \max\{1, \lceil \rho \min(d_{\text{out}}, d_{\text{in}}) \rceil\}$ and let U_k and V_k contain the first k left and right singular vectors.

We quantify the capacity-normalized task-update energy captured by these subspaces using Direction-Band Alignment (DBA):

$$\text{DBA}_t(M; k) = \frac{d_{\text{out}} \left\| U_k^\top \Delta W_t \right\|_F^2 + d_{\text{in}} \left\| \Delta W_t V_k \right\|_F^2}{2k \left\| \Delta W_t \right\|_F^2}. \tag{8}$$

The two terms measure captured energy in the output and input subspaces, respectively. Their capacity factors make the two sides comparable, and normalization by $\|\Delta W_t\|_F^2$ makes the diagnostic invariant to the scale of the task update. Singular values determine which directions enter the band but do not otherwise weight the score. DBA therefore evaluates directional structure rather than spectral strength.

For every matrix-valued layer $\ell \in \mathcal{L}$, we set $k_\ell = \max\left\{1, \left\lceil \rho \min\left(d_{\text{out}}^{(\ell)}, d_{\text{in}}^{(\ell)}\right) \right\rceil\right\}$. The model-level score for task t is the unweighted layer average

$$\text{DBA}_t^{\text{model}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{DBA}_t\left(M^{(\ell)}; k_\ell\right). \tag{9}$$

We average this score across tasks when reporting one aggregate value and use $\rho = 0.1$ by default. Unlike the spectrum-dependent effective rank used by the subspace alignment ratio (Marczak et al., 2025), this fixed band keeps directional capacity constant across methods.

Figure 1 shows that RSS-Merge achieves the highest DBA on all eight tasks, whereas TSV-Merge, DC-Merge, and CART remain below TA. Under the same leading-band capacity, the dominant singular subspaces of RSS-Merge therefore capture more task-update energy than those of the compared baselines. Because spectrum calibration preserves the singular directions and their ordering, the final update and Δ_{rel} have identical DBA. The observed advantage thus supports the intended structural effect of relevance construction. DBA is a diagnostic of directional structure rather than a proxy for downstream accuracy; downstream utility is assessed separately by task performance and controlled ablations.

RSS-Merge continues to achieve the highest DBA among the compared methods across the tested diagnostic bandwidths; see Appendix C.1 for the full sensitivity analysis. We further verify that the task-wise DBA advantage of RSS-Merge persists on ViT-B/16 and ViT-L/14; see Appendix C.3.

4.3.2 Contributions of Relevance Construction and Spectrum Calibration

We analyze whether relevance construction provides a consistent benefit both before and after spectrum calibration. Using TA as the unmodified aggregation baseline, we compare TA with the

relevance-constructed update when calibration is absent, and compare their calibrated counterparts when calibration is applied. Table 3 reports these two matched comparisons, which hold the calibration state fixed and isolate the contribution of relevance construction.

Table 3: Ablation of relevance construction and spectrum calibration on ViT-B/32 across task-set sizes. We report average absolute accuracy (%). Each row uses the same validation protocol to select the global rescaling coefficient α.

Variant	8 tasks	14 tasks	20 tasks
TA	69.12	65.24	60.47
TA + spectrum calibration	86.30	80.09	75.22
Relevance construction only	75.20	70.47	66.82
RSS-Merge	**88.76**	**84.14**	**81.23**

Holding the calibration state fixed, relevance construction consistently improves performance across all task-set sizes. Without spectrum calibration, it improves over TA by 6.08, 5.23, and 6.35 percentage points in the 8-task, 14-task, and 20-task settings, respectively. When both variants use spectrum calibration, the complete RSS-Merge further improves over calibrated TA by 2.46, 4.05, and 6.01 points. Thus, the benefit of relevance construction persists after calibration and grows with task count in the calibrated comparison.

4.3.3 Role of the Complementary Branch and Bilateral Projection

We first test whether the complementary average-update branch is necessary by retaining only the residual-guided update Δ_{res}. Starting from this residual-only variant, the no-projection variant adds the average update $\overline{\Delta W}$ directly, while the left-projection variant uses $(I - U_{\mathrm{res}}U_{\mathrm{res}}^{\top})\overline{\Delta W}$. The complete bilateral construction additionally removes overlap in the right subspace, as defined in Equation 2. All other components and hyperparameters are kept unchanged.

Table 4: Ablation of the complementary average-update branch and its projection on ViT-B/16 with 8 and 14 tasks. We report average absolute accuracy (%).

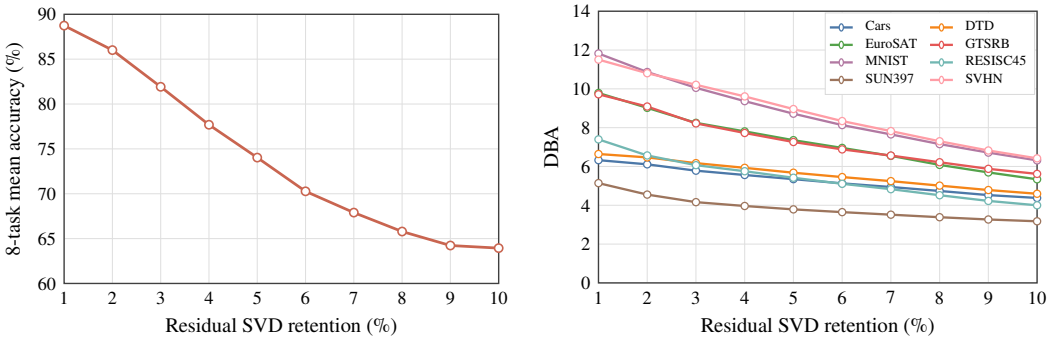
Variant	8 tasks	14 tasks
Residual-guided branch only	74.36	77.05
+ Unprojected average-update branch	90.75	85.27
+ Left-projected average-update branch	90.44	85.44
+ Bilaterally projected average-update branch	**91.96**	**87.31**

As shown in Table 4, adding even the unprojected average-update branch to the residual-guided update improves average accuracy by 16.39 and 8.22 percentage points in the 8-task and 14-task settings, respectively. Among the tested ways of adding this branch, bilateral projection performs best: it improves over no projection by 1.21 and 2.04 points and over left projection by 1.52 and 1.87 points. These comparisons show that the residual branch alone omits useful average-update information and that bilateral separation yields the strongest result among the tested constructions.

4.3.4 Selection of the Residual Retention Ratio

The relevance-construction stage retains the leading p fraction of singular components from each centered task update. We examine the choice of $p = 1\%$ by varying p from 1% to 10% on the ViT-B/32 eight-task benchmark and measuring both average absolute accuracy and DBA. Unlike the diagnostic bandwidth ρ, p changes the merged update itself; we keep $\rho = 0.1$ when comparing DBA across retention ratios. The global rescaling coefficient α follows the validation protocol described above.

Figure 2a shows that the merged model's average accuracy decreases as more singular components are retained from the centered task updates. Figure 2b shows a corresponding decrease in the DBA



(a) Effect of residual retention on accuracy.

(b) Effect of residual retention on DBA.

Figure 2: Residual retention in RSS-Merge on ViT-B/32 with eight tasks. Both panels vary the percentage of leading singular components retained from each centered task update. (a) Absolute accuracy of the resulting merged model, averaged over the eight tasks. (b) DBA of the resulting merged update for each task at a fixed diagnostic bandwidth $\rho = 0.1$.

of the merged update for every task, with the diagnostic bandwidth held fixed. These observations support retaining a small fraction of residual components in this setting: among the tested ratios, $p = 1\%$ yields both the highest average accuracy and the highest mean DBA. This evidence supports our default choice within the tested range, without implying that DBA alone determines downstream accuracy.

4.3.5 Robustness to the Residual Weight

The coefficient λ controls the relative spectral strength of the residual-guided branch and the complementary average-update branch in the relevance-constructed update. We therefore examine whether the performance of RSS-Merge depends on a narrowly tuned balance between these two sources of structure. To this end, we vary λ from 2.5 to 3.5 on ViT-B/32 with eight tasks. As shown in Figure 3, average accuracy stays near 88.5%–88.8%, with a variation of only about 0.3 percentage points. Every tested weight remains above the strongest compared merging baseline in this setting. This stable plateau shows that strong performance is sustained across the tested weights, supporting robustness to the balance between the two branches.

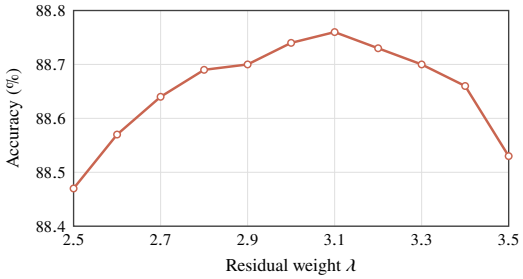


Figure 3: Average accuracy on the 8-task ViT-B/32 setting under different residual weights.

5 Conclusion

We study model merging from the perspective that a single merged update generally cannot preserve the full directional geometry of multiple distinct task updates simultaneously. This motivates prioritizing task-relevant structure within a dominant low-rank representation of the final merged update, while separately controlling how strongly the resulting directions are expressed through the singular spectrum. Based on this principle, we propose RSS-Merge, which constructs a relevance-oriented directional structure from salient task-relative residuals and the complementary component of the mean update, followed by order-preserving spectrum calibration that smooths spectral disparities without altering the constructed directions or their ordering. Extensive experiments on vision and vision-language benchmarks consistently demonstrate the effectiveness of RSS-Merge: it achieves the best average accuracy across all nine vision settings, as well as the highest unseen-task average and second-highest seen-task average on MM-MergeBench among the compared merging methods. These results support treating directional relevance and spectral weighting as distinct but complementary design axes for model merging.

AI USE STATEMENT

We used ChatGPT to assist with language polishing and improve the clarity and readability of the manuscript. The authors reviewed and revised the resulting text and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

We will release our code as open source, including the model-merging implementation, evaluation scripts, and experimental configurations, to facilitate independent reproduction of our results. Implementation details and hyperparameter settings will be documented in the paper, supplementary material, and accompanying code repository.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 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.

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 *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Jiho Choi, Donggyun Kim, Chanhyuk Lee, and Seunghoon Hong. Revisiting weight averaging for model merging, 2024.

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

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

Adam Coates, Honglak Lee, and Andrew Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks*, pp. 2921–2926, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim K Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. In *Advances in Neural Information Processing Systems*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pp. 417–434, 2020.

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.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Xisen Jin, Xiang Ren, Daniel Preoţiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations*, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations*, 2023.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartlomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, 2022.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition*, pp. 947–952, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

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 *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162, 2022.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the knots. In *International Conference on Learning Representations*, 2025.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology, 2018.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2Words: Automatic mobile UI summarization with multimodal learning. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 498–510, 2021.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak 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 *Proceedings of the 39th International Conference on Machine Learning*, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Fanhu Zeng, Haiyang Guo, Fei Zhu, Li Shen, and Hao Tang. Robustmerge: Parameter-efficient model merging for mllms with direction robustness. In *Advances in Neural Information Processing Systems*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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, 2026.

A Implementation Details

Choice of checkpoints. For the fully fine-tuned vision experiments in Table 1, we use the task-specific checkpoints released by TSV-M (Gargiulo et al., 2025). For the MM-MergeBench experiments in Table 2, we use the LLaVA-v1.5-7B LoRA checkpoints released by RobustMerge (Zeng et al., 2025).

Choice of hyperparameters. For the vision experiments, we set the residual retention ratio to $p = 1\%$, the residual weight to $\lambda = 3$, and the smoothing coefficient to $\gamma = 0.99$. We select the global rescaling coefficient α on the held-out validation split from $\{2.0, 2.1, \ldots, 4.0\}$, independently for each variant and hyperparameter setting. For MM-MergeBench, we set $p = 1\%$ and $\gamma = 0.99$ and jointly sweep $\lambda \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ and $\alpha \in \{2.0, 2.1, \ldots, 4.0\}$.

B Dataset Description

The vision benchmark follows the fully fine-tuned CLIP protocol of Wang et al. (2024). Its 8-task, 14-task, and 20-task configurations are nested: the 14-task setting augments the 8-task setting with six datasets, and the 20-task setting adds a further six datasets.

The 8-task setting comprises the following datasets.

- **Stanford Cars (Cars)** (Krause et al., 2013): fine-grained recognition of car make, model, and production year.
- **Describable Textures Dataset (DTD)** (Cimpoi et al., 2014): recognition of natural and human-made textures from describable visual attributes.
- **EuroSAT (EuroSAT)** (Helber et al., 2019): land-use and land-cover classification from Sentinel-2 satellite images.
- **German Traffic Sign Recognition Benchmark (GTSRB)** (Stallkamp et al., 2011): recognition of German traffic signs captured under varying conditions.
- **MNIST (MNIST)** (LeCun et al., 1998): classification of grayscale handwritten digits.
- **NWPU-RESISC45 (RESISC45)** (Cheng et al., 2017): scene classification for aerial imagery.
- **Street View House Numbers (SVHN)** (Netzer et al., 2011): recognition of digits cropped from house numbers in Google Street View images.
- **SUN397 (SUN397)** (Xiao et al., 2010): classification of indoor, outdoor, and natural scenes.

The 14-task setting includes the eight datasets above and the following six additional datasets.

- **STL-10 (STL10)** (Coates et al., 2011): classification of natural images from ten object classes.
- **Oxford-IIIT Pet (OxfordIIITPet)** (Parkhi et al., 2012): fine-grained classification of 37 cat and dog breeds.
- **Oxford 102 Flowers (Flowers102)** (Nilsback & Zisserman, 2008): fine-grained classification of 102 flower categories across varying scales, poses, and lighting conditions.
- **CIFAR-100 (CIFAR100)** (Krizhevsky, 2009): classification of low-resolution natural images into 100 object classes.
- **PatchCamelyon (PCAM)** (Veeling et al., 2018): detection of metastatic tissue in histopathology patches.
- **Facial Expression Recognition 2013 (FER2013)** (Goodfellow et al., 2015): classification of grayscale face images into seven expression categories.

The 20-task setting includes all 14 datasets above and the following six additional datasets.

- **CIFAR-10 (CIFAR10)** (Krizhevsky, 2009): classification of low-resolution natural images into ten object classes.
- **Food-101 (Food101)** (Bossard et al., 2014): classification of food photographs into 101 dish categories.

- **Fashion-MNIST (FashionMNIST)** (Xiao et al., 2017): classification of grayscale clothing and accessory images into ten categories.
- **Rendered SST-2 (RenderedSST2)** (Socher et al., 2013; Ilharco et al., 2023): binary sentiment classification on SST-2 sentences rendered as images.
- **EMNIST Digits (EMNIST)** (Cohen et al., 2017): handwritten-digit classification using the digit subset of EMNIST.
- **Kuzushiji-MNIST (KMNIST)** (Clanuwat et al., 2018): classification of ten cursive Japanese Kuzushiji character classes.

B.1 MM-MERGEBENCH

MM-MergeBench (Zeng et al., 2025) contains eight seen tasks whose task-specific checkpoints are merged and four unseen tasks used only to evaluate transfer beyond the merged task set.

The seen-task group comprises the following eight datasets.

- **ScienceQA (SciQA)** (Lu et al., 2022): multiple-choice science question answering with image or diagram context.
- **ImageNet (Image)** (Deng et al., 2009): large-scale object recognition, formulated as generative image classification in MM-MergeBench.
- **VQAv2 (VQA)** (Goyal et al., 2017): open-ended question answering about natural images.
- **REC-COCO (REC)** (Mao et al., 2016): localization of the image region specified by a natural-language expression.
- **OCR-VQA (OCR)** (Mishra et al., 2019): visual question answering over text embedded in images.
- **VizWiz-Caption (VizWiz)** (Gurari et al., 2020): captioning of images captured by blind or visually impaired users.
- **Flickr30k (Flickr)** (Young et al., 2014): captioning of real-world photographs.
- **IconQA (IconQA)** (Lu et al., 2021): visual-language reasoning over abstract diagrams and icon-based scenes.

The unseen-task group comprises the following four datasets.

- **A-OKVQA (AVQA)** (Schwenk et al., 2022): visual question answering that draws on commonsense and world knowledge beyond the image.
- **ImageNet-R (Image-R)** (Hendrycks et al., 2021): object recognition on artistic renditions of ImageNet classes.
- **Screen2Words (S2W)** (Wang et al., 2021): generation of concise natural-language summaries for mobile user-interface screens.
- **TabMWP (TAbMWP)** (Lu et al., 2023): mathematical reasoning over semi-structured tables and associated word problems.

C ADDITIONAL STRUCTURAL DETAILS

C.1 SENSITIVITY TO THE DIAGNOSTIC BANDWIDTH

To assess whether the directional comparison depends on the diagnostic bandwidth, Figure 4 reports the relative DBA change (%) with respect to TA for $\rho \in \{0.05, 0.10, 0.15, 0.20\}$. RSS-Merge achieves the highest DBA among the compared methods at every tested ratio, while the other methods remain below it throughout. The observed directional relevance advantage therefore persists across the tested bandwidths.

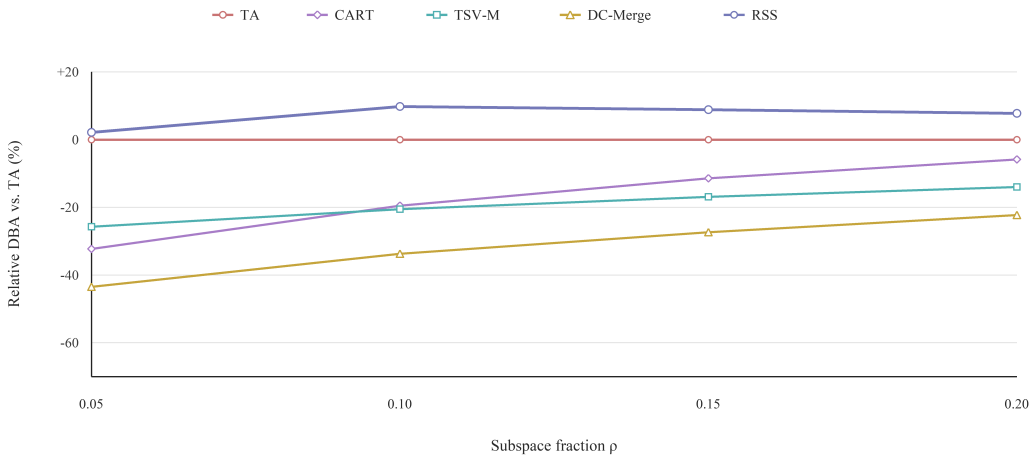


Figure 4: Relative DBA change (%) versus TA across ρ on ViT-B/32 with eight tasks.

C.2 Validity of Averaging Left- and Right-Sided Alignment

DBA averages two complementary quantities: the task-update energy captured by the leading left singular subspace and that captured by the leading right singular subspace. To expose possible asymmetry between the two sides, we define

$$\mathrm{DBA}_t^{(L)}(M;k) = \frac{d_{\mathrm{out}} \left\| U_k^\top \Delta W_t \right\|_F^2}{k \left\| \Delta W_t \right\|_F^2}, \qquad \mathrm{DBA}_t^{(R)}(M;k) = \frac{d_{\mathrm{in}} \left\| \Delta W_t V_k \right\|_F^2}{k \left\| \Delta W_t \right\|_F^2}. \tag{10}$$

The aggregate metric in Equation 8 satisfies $\mathrm{DBA}_t = \frac{1}{2}(\mathrm{DBA}_t^{(L)} + \mathrm{DBA}_t^{(R)})$. Reporting the two components separately reveals whether an aggregate DBA difference is supported by both matrix subspaces or is driven primarily by one side. This decomposition should be interpreted as a diagnostic of left–right balance; the main metric intentionally summarizes the two marginal notions of relevance rather than measuring projection onto paired rank-one singular components.

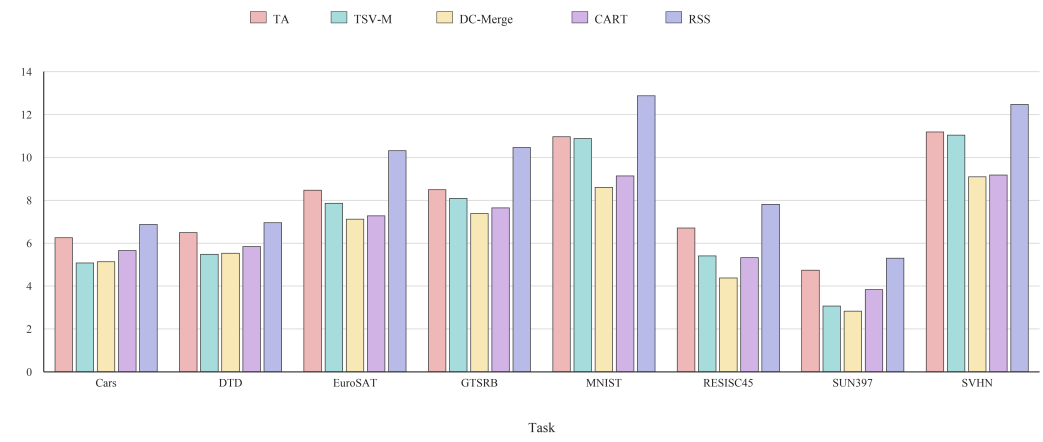


Figure 5: Task-wise left-sided DBA on the ViT-B/32 eight-task benchmark with $\rho = 0.1$.

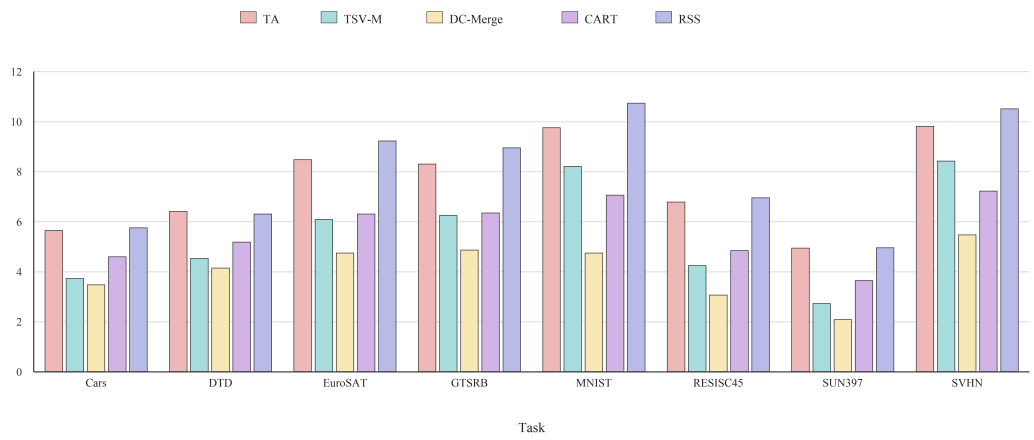


Figure 6: Task-wise right-sided DBA on the ViT-B/32 eight-task benchmark with $\rho = 0.1$.

Figures 5 and 6 verify that the aggregate DBA comparison is supported by both matrix sides rather than being dominated by one projection. RSS-Merge obtains the highest left-sided DBA on all eight tasks and the highest right-sided DBA on seven tasks; DTD is the only right-sided exception, where TA is marginally higher. The broadly consistent ranking across the two sides supports their average as a balanced model-level diagnostic, while their task-dependent differences show why retaining both components is preferable to using either side alone.

C.3 Robustness Across Model Scales

We further evaluate the averaged DBA on larger CLIP vision backbones to determine whether the diagnostic remains informative beyond ViT-B/32. Figures 7 and 8 report task-wise DBA for ViT-B/16 and ViT-L/14, respectively, under the same eight-task protocol.

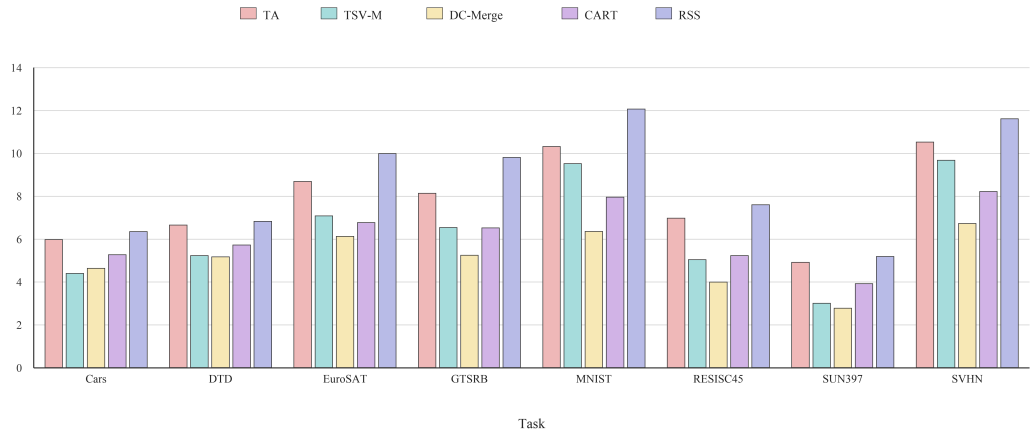


Figure 7: Task-wise DBA on the ViT-B/16 eight-task benchmark with $\rho = 0.1$.

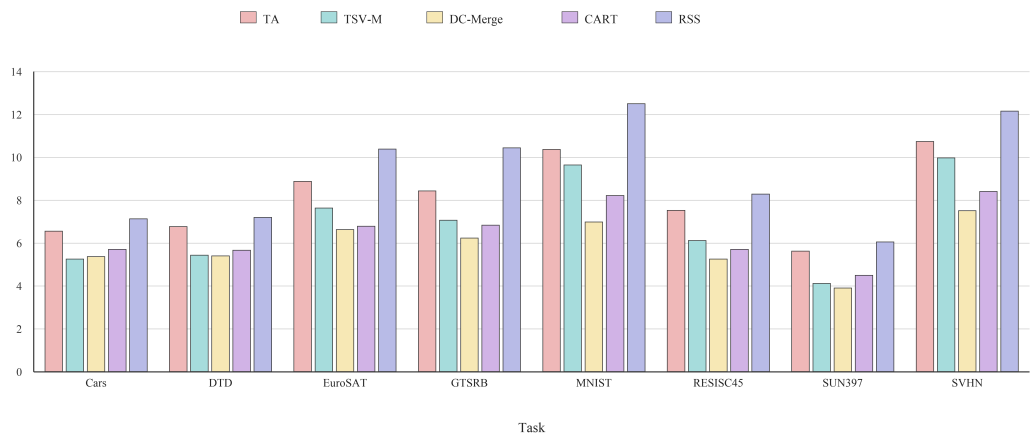


Figure 8: Task-wise DBA on the ViT-L/14 eight-task benchmark with $\rho = 0.1$.

Across both larger backbones, RSS-Merge achieves the highest DBA on all eight tasks. Together with the ViT-B/32 results in Figure 1, these consistent task-wise rankings show that the structural advantage of RSS-Merge persists across backbone scales rather than being specific to ViT-B/32.