WHAT IS A MERGE WORTH? DATA AND TRAINING BUDGETS IN MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging can produce a checkpoint for immediate deployment or an initialization for further adaptation. These uses raise different questions: a merged model's deployment advantage may change when the pretrained base is also allowed to train, and single-budget evaluations cannot separate the value of initialization from access to more data or optimization. We compare fixed TSV and pretrained CLIP initializations across eight classification tasks, using matched adaptation and paired seeds to measure changes in their performance gap along data and update axes that share a common endpoint. On ViT-B/32, increasing the fitting pool reduces the gap more than extending training on the full pool; on ViT-B/16, the data contrast points in the same direction but remains uncertain. TSV retains an advantage at the shared endpoint, while supervised and magnitude controls show that the result is not limited to distillation or explained by update norm alone. The findings suggest that practitioners compare merged and pretrained starts across adaptation budgets and interpret merge value within the evaluated access regime, rather than as a universal trade-off or end-to-end saving.

1 INTRODUCTION

Model merging has emerged as a powerful technique to combine multiple task specialists into a single checkpoint without requiring additional training. While a merged model is often evaluated based on its immediate zero-shot deployment performance, its potential as a starting point for further downstream adaptation is equally important but less understood. This distinction is highly relevant when task specialists are already available. Practitioners can easily reuse the initial training investment captured by merging techniques, but they still face the practical challenge of allocating limited computational resources and data to finalize the adaptation for their specific needs.

Existing literature provides various methods for combining related models, yet evaluations typically rely on a single adaptation budget. This approach obscures the underlying mechanics of why a merged initialization is beneficial. Specifically, it remains unclear whether a merged starting point achieves better performance because it uses fitting data more efficiently, because it requires fewer optimizer updates, or a combination of both. Increasing the available fitting data expands the diversity of examples the model sees, whereas increasing optimizer updates simply extends the optimization trajectory over an existing pool of images. Evaluating these factors simultaneously confounds their individual effects. Therefore, we ask a precise question: *how does the performance gap between a merged initialization and a pretrained base model change when we independently scale either the fitting data or the optimizer updates?*

To answer this, we conduct a controlled empirical study using fixed TSV-Merge initializations and pretrained CLIP vision encoders, specifically ViT-B/32 and ViT-B/16, evaluated across eight standard image classification tasks. For every experimental setting, both the merged model and the pretrained base utilize identical fitting images, specialist predictions, and adaptation recipes across paired optimization seeds. We design two distinct intervention paths that converge on a shared, long-run endpoint with full data access. The first intervention expands the fitting pool while holding the number of updates constant, and the second extends the training duration while holding the data pool constant. We do not treat these interventions as strictly cost-equivalent. Instead, we measure the extent to which each resource reduces the initialization advantage held by the merged model.

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.

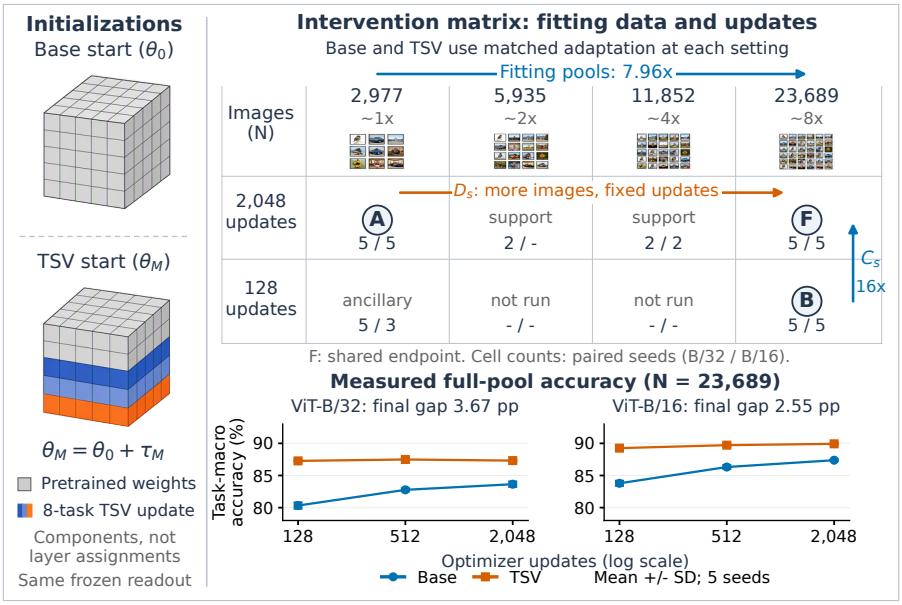

Figure 1: **Initializations, interventions, and measured outcomes:** Left: the pretrained base and a fixed TSV initialization from eight specialists. Weight blocks and image tiles are schematic, not literal layer assignments or benchmark samples. Upper right: the data path $A \to F$ holds updates at 2,048; the update path $B \to F$ holds fitting images at 23,689. Both starts use matched adaptation. Cell counts report paired seeds for ViT-B/32 and ViT-B/16; supporting and ancillary settings are distinguished from the five-seed primary endpoints. Lower right: full-pool task-macro accuracy on the reused test collection, with means $\pm$ sample SD over five seeds. Segments connect measured checkpoints, not extrapolated convergence curves.

As illustrated in Figure 1, the pretrained base does improve with additional full-pool training, but it consistently trails the TSV initialization even at the largest measured budget.

Our primary contributions are detailed as follows: ❶ **Separating resource interventions through a common-endpoint design.** We explicitly quantify how the initialization gap shrinks when scaling either the fitting pool size or the update budget. This contrast isolates the impact of each resource for our measured settings rather than proposing a universal exchange rate between data and compute. ❷ **Demonstrating backbone-dependent gap reductions.** The relative benefit of data versus compute varies by model architecture. On ViT-B/32, increasing data reduces the initialization gap significantly more than increasing updates, showing a difference of 6.96 percentage points. While ViT-B/16 exhibits a similar directional trend, the variance across our five paired seeds leaves the contrast unresolved for that backbone. Nevertheless, the merged initialization maintains a measurable advantage at the shared endpoint for both models. ❸ **Constraining explanations via rigorous controls.** We verify that the merged initialization gap persists under standard supervised training, ensuring the effect is not an artifact of knowledge distillation. Furthermore, norm-matched random perturbations and shuffled updates fail to reproduce the TSV advantage, confirming that the benefit stems from the structured knowledge within the merged weights rather than mere update magnitude.

This study characterizes the adaptation dynamics when specialist models are treated as existing assets. We do not account for the original computational cost of training these specialists, and all target accuracies are reported on an already-exposed test collection. Consequently, our findings reflect the measured adaptation benefits under specific data and update budgets rather than claiming end-to-end efficiency guarantees or proposing a universal scaling law for model merging.

2 Related Work

Combining models with shared ancestry: Many merging methods combine models fine-tuned from a shared pretrained checkpoint. Weight averaging combines their parameters directly (Wortsman et al., 2022a), while task arithmetic represents each specialist as a change from the common

initialization (Ilharco et al., 2023). TIES modifies this combination to address sign interference (Yadav et al., 2023), and AdaMerging learns task- or layer-wise combination coefficients (Yang et al., 2024). These methods differ in how they form a merged checkpoint, but our question is how its advantage changes when the pretrained base receives more adaptation data or updates. We therefore hold the main comparison to one fixed merging method; a separate initialization sweep includes averaging, task arithmetic, and TIES.

Subspace structure and downstream adaptation: Several approaches use subspace structure to reduce interference between task updates. TSV uses singular-vector structure (Gargiulo et al., 2025); KnOTS aligns updates in shared SVD coordinates, particularly for low-rank updates (Stoica et al., 2025); and ESM uses feature shifts to identify merging subspaces (Li et al., 2026). These methods develop different ways to construct a merge. We use TSV as a fixed initializer rather than proposing or ranking merging operators, and measure its downstream behavior across explicit data and update budgets. MetaMerging also targets downstream adaptation: it learns merge coefficients through simulated adapter updates (Chen et al., 2026). Unlike that approach, our study keeps the merge and adaptation recipe fixed to isolate how the initialization gap changes across the selected resource settings.

Initialization, transfer, and distillation: Pretrained initialization can affect both feature reuse and the optimization landscape (Neyshabur et al., 2020). Knowledge distillation transfers behavior through teacher predictions (Hinton et al., 2015), while FS-Merge learns a foldable supernetwork through feature reconstruction (Kinderman et al., 2025). Our full-vision distillation is a conventional adaptation baseline, not FS-Merge. The primary comparison asks a different question from these methods: under the same teacher signal and training recipe, how does the merged initialization's advantage change as fitting images or optimizer updates increase?

Geometry and retained capability: Permutation alignment and interpolation repair help explain when parameter combinations preserve network behavior (Ainsworth et al., 2023; Jordan et al., 2023). Our entry-shuffling control is not a function-preserving neuron permutation; it tests whether tensor norms and entry distributions alone account for the gain. Separately, WiSE-FT shows that interpolating fine-tuned weights with a zero-shot model can improve robustness (Wortsman et al., 2022b). Our retention experiment follows this established principle and measures a specific target-accuracy trade-off. The additional interpolation sweep and its results are presented in Appendix E; interpolation itself is not a new method here.

3 METHODS

3.1 MODEL SCOPE AND FIXED INITIALIZATIONS

Common-base vision merging: For each backbone, a pretrained vision encoder θ_0 and eight public specialists $\theta_k = \theta_0 + \tau_k$ share the same parameterization. Here, τ_k is the task-specific change in vision weights and $K = 8$. The tasks are Cars, DTD, EuroSAT, GTSRB, MNIST, RESISC45, SUN397, and SVHN, with 758 classes in total. Evaluation assumes the task identity is known, so each image is classified using that task's class vocabulary. This is not arbitrary vision–language model fusion or task-agnostic classification.

Network and trainable scope: Both starts use the same CLIP architecture and frozen readout. The vision encoders have 12 Transformer blocks of width 768. Their 32- or 16-pixel patches produce 49 or 196 patch tokens from a 224-pixel image, plus a class token. Let $v_\theta(x) \in \mathbb{R}^{768}$ be the pooled class-token representation. Prediction uses the frozen visual projection $P_0 \in \mathbb{R}^{512\times768}$, normalized text-derived class vectors as the rows of $H_k \in \mathbb{R}^{|\mathcal{C}_k|\times512}$ for task vocabulary $\mathcal{C}_k$, and pretrained logit scale s_0:

$$p_\theta(\cdot \mid x, k) = \mathrm{softmax}\left(s_0 H_k \frac{P_0 v_\theta(x)}{\|P_0 v_\theta(x)\|_2} \right). \tag{1}$$

We merge and adapt only the vision encoder, including its embeddings and layer norms. The visual projection, text tower and its projection, cached class heads, and logit scale remain at their pretrained values. Gradients pass through the fixed readout, but only vision-encoder parameters are updated.

The fixed merged initializer: We write the TSV start as

$$\theta_M = \theta_0 + \tau_M, \qquad \tau_M = \mathcal{M}_{\mathrm{TSV}}(\tau_1, \ldots, \tau_K). \tag{2}$$

For an $m \times n$ matrix, TSV retains the leading $r = \lfloor \min(m, n)/K \rfloor$ singular components from each task update. It concatenates the retained components, replaces the left and right factors by their polar factors, and recomposes the merged matrix. Here, $\mathrm{polar}(X) = PQ^\top$ for a thin SVD $X = P\Sigma Q^\top$. Nonmatrix updates are averaged; attention QKV is packed before this operation. The global merge scale is 1. Each backbone has one fixed TSV checkpoint, reused across all primary pool sizes, update budgets, and optimization seeds; neither the merging rule nor its scale is retuned for a resource setting. Both backbones use public `tanganke` specialists, identical 224-pixel CLIP preprocessing, and the same prompt templates, with text heads built separately for each backbone.

3.2 Matched adaptation and data access

Matched adaptation: Let $\mathcal{T}_s(\theta; N, U)$ denote adaptation from θ using an N-image fitting pool, U optimizer updates, and optimization seed s. We compare $\mathcal{T}_s(\theta_0; N, U)$ with $\mathcal{T}_s(\theta_M; N, U)$ at every experimental resource setting. The base receives no expert-derived weight update; both starts access the same specialists through predictions on the fitting pool. Matching the recipe does not require identical learned parameter changes.

The primary objective is task-balanced teacher-to-student KL divergence. Writing $\mathcal{D}_k(N)$ for the empirical fitting distribution of task k,

$$\mathcal{L}_{\mathrm{KD}}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{x \sim \mathcal{D}_k(N)} \, , \qquad K = 8. \tag{3}$$

The temperature is 1. Each update averages two batches of 16 images per task, or 256 student-image presentations in total. We use AdamW with a constant learning rate of 10^{-5}, zero weight decay, and gradient clipping at norm 1. Within each condition and seed, the two initializations share the fitting subset, teachers, and sampled batch sequence. The 128-, 512-, and 2,048-update measurements are saved checkpoints of one continuing run, not separately selected stopping points. The supervised control replaces KL with cross-entropy on ground-truth labels while retaining the same update and sampling recipe.

Fitting and selection roles: The primary fitting pool contains 23,689 TRAIN images, up to 32 per class. It includes the original 6,064-image fitting set used in the supporting controls. Class-stratified subsampling gives pools of 2,977, 5,935, 11,852, and 23,689 images; all classes are represented at every level. Here, pool size means *distinct images available for sampling*, not a guarantee that every image is observed at every update budget. A 3,032-image selection role and two 3,032-image development roles are disjoint from the fitting pool and each other. Public specialists may have seen any of these TRAIN examples during their original fine-tuning. Distillation does not use class labels in its loss, but labels are used to construct the stratified fitting pools. We therefore do not describe its data access as fully label-free.

3.3 Resource interventions and estimands

Let $A_{\mathrm{TSV},s}(N, U)$ and $A_{\mathrm{base},s}(N, U)$ be task-macro accuracies, in percent, after U updates using a pool of N images and seed s. For a fixed backbone, define the initialization gap

$$G_s(N, U) = A_{\mathrm{TSV},s}(N, U) - A_{\mathrm{base},s}(N, U) \, . \tag{4}$$

A positive gap favors TSV initialization. Set $N_\ell = 2{,}977$, $N_h = 23{,}689$, $U_\ell = 128$, and $U_h = 2{,}048$. We measure the reduction in this gap under two interventions:

$$C_s = G_s(N_h, U_\ell) - G_s(N_h, U_h) \, , \qquad \text{more updates at fixed pool,} \tag{5}$$

$$D_s = G_s(N_\ell, U_h) - G_s(N_h, U_h) \, , \qquad \text{more images at fixed updates.} \tag{6}$$

Each quantity is a change in the *initialization gap*, not the accuracy gain of either model alone. The primary contrast is $\Delta_s = D_s - C_s$. Both reductions use the same full-pool, 2,048-update endpoint within seed. Consequently,

$$\Delta_s = G_s(N_\ell, U_h) - G_s(N_h, U_\ell).$$

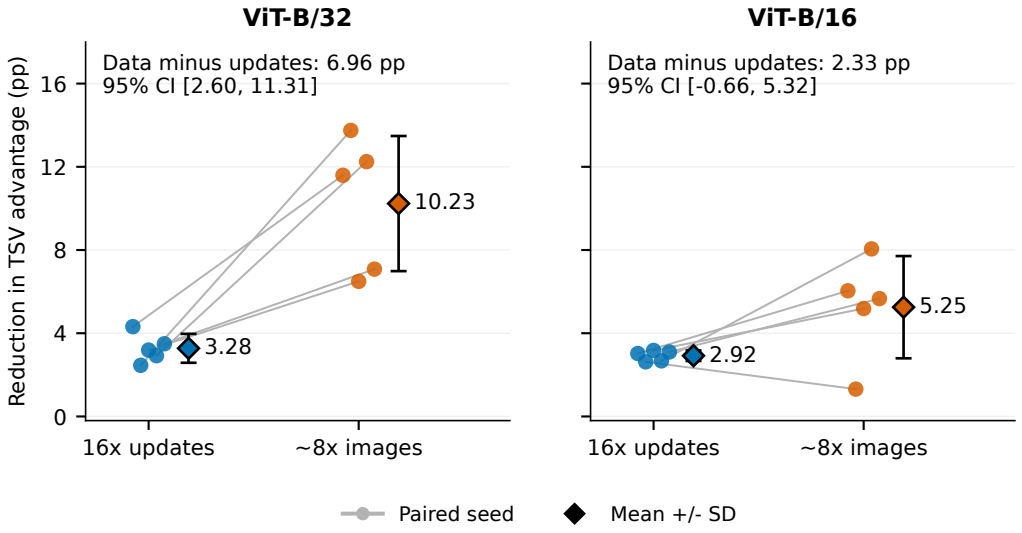

Figure 2: **Gap reduction under two resource interventions:** Each line connects one seed's reduction from $16\times$ more updates (C_s) to its reduction from approximately $8\times$ more fitting images (D_s). Diamonds and error bars show means and sample standard deviations across five seeds. Insets report the paired difference $D_s - C_s$ and its 95% $t(4)$ interval. Data yields the larger reduction on ViT-B/32; the ViT-B/16 contrast is smaller and its interval includes zero. Both panels use the same vertical scale.

A positive value means that the specified pool expansion reduces the gap more than the specified update extension. The interventions increase updates 16-fold and available fitting images approximately 7.96-fold. At fixed U, pool expansion changes the available image diversity but preserves the $256U$ student-image presentations.

This design separates *two finite interventions, not their economic cost*. They start from different resource settings, and a sixteen-fold update increase is not equivalent to an eight-fold data increase. "Compute" refers here to student optimizer updates within a backbone. It excludes expert training, merging, teacher-logit preparation, evaluation, and hyperparameter search. Equal update counts are not equal FLOPs across backbones. The comparison asks how much each intervention reduces the initialization gap; it does not estimate a universal exchange rate or an end-to-end saving.

3.4 Evaluation and statistical scope

We compute accuracy within each task and average equally over the eight tasks. Unless stated otherwise, all accuracies use the full, already-exposed 87,433-image provider test collection. Earlier test measurements informed research decisions, so these are exploratory measurements on a reused test set, not an untouched confirmation set.

The primary comparison uses seeds 117–121 on each backbone. We report sample standard deviations across optimization seeds and 95% paired Student t intervals for seed-level contrasts. For a five-seed contrast X_s, the interval is $\bar{X} \pm t_{0.975,4}\,\mathrm{SD}(X_s)/\sqrt{5}$. These intervals condition on the tasks, experts, fitting subsets, and test examples; they do not quantify uncertainty over new tasks, experts, or fitting pools. Overlapping ViT-B/16 campaign records are checked for agreement and counted once. Supporting controls use two or three seeds, as stated locally, and are not pooled with the primary estimates. All inferential comparisons are exploratory and unadjusted for multiple analyses.

4 Results: data and update budgets

The data intervention removes more of the ViT-B/32 gap: At 2,048 updates, increasing the fitting pool reduces the initialization gap from 13.91 to 3.67 points (Table 1). The reduction is

Table 1: **Endpoints of the primary comparison:** Mean task-macro accuracy (%) on the full reused test collection, with five seeds in every row. Gap is TSV minus base, reported as mean ± sample SD in percentage points and computed before rounding. The final row for each backbone is the common endpoint of the data and update interventions. Stronger/lighter green marks the best/second distinct displayed mean within each backbone and accuracy column; ties share a rank. Here and below, shading describes point estimates, not statistical significance.

Backbone	Fit images	Updates	Base	TSV	Gap (pp)
ViT-B/32	2,977	2,048	70.39	84.30	13.91 ± 3.58
ViT-B/32	23,689	128	80.32	87.27	6.95 ± 0.49
ViT-B/32	23,689	2,048	83.64	87.32	3.67 ± 0.47
ViT-B/16	2,977	2,048	77.54	85.34	7.80 ± 2.27
ViT-B/16	23,689	128	83.79	89.25	5.47 ± 0.22
ViT-B/16	23,689	2,048	87.38	89.92	2.55 ± 0.25

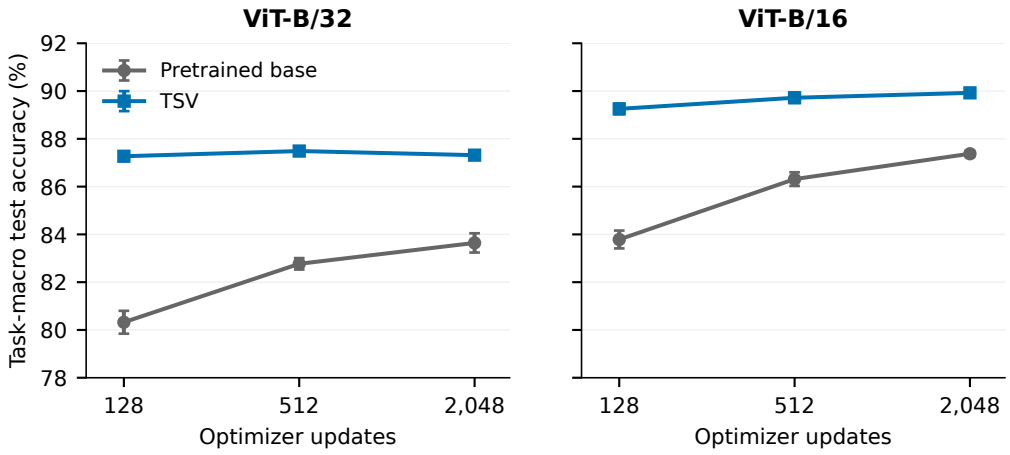

Figure 3: **The base improves with more updates, but still trails:** Full-pool distillation on both backbones with 23,689 available fitting images and five seeds. Markers show mean task-macro accuracy on the full reused test collection; error bars show sample SD. The common axis scales make both the initialization gap and backbone difference visible. These curves show finite-budget performance, not proof of convergence.

10.23 ± 3.25 points. By comparison, increasing updates on the full pool reduces the gap from 6.95 to 3.67 points, a reduction of 3.28 ± 0.70 points. The paired difference is 6.96 points, with 95% interval $[2.60, 11.31]$ and two-sided $p = 0.011$. *The data intervention produces the larger reduction in all five seeds* (Figure 2).

The absolute accuracies clarify what is changing. On the data axis, base accuracy rises from 70.39 to 83.64, while TSV accuracy rises from 84.30 to 87.32. The smaller gap is therefore not an artifact of degrading the merged model. On the update axis, TSV is already near 87.3 at 128 updates, while the base improves from 80.32 to 83.64. Figure 3 also shows the intermediate 512-update measurements. Additional updates help, but do not erase the gap over the measured range.

A common starting point gives the same contrast: The ViT-B/32 campaign covers both fitting pools at all three update budgets. The 2,977-image, 128-update setting has a mean gap of 15.68 points. Starting from that shared setting, increasing data reduces the gap by 8.73 points, while increasing updates reduces it by 1.77 points. The difference is again 6.96 points, identically within every seed. This is an alternative expression of the same contrast: subtracting both reductions from the same starting gap again leaves Δ_s. It is not an independent replication.

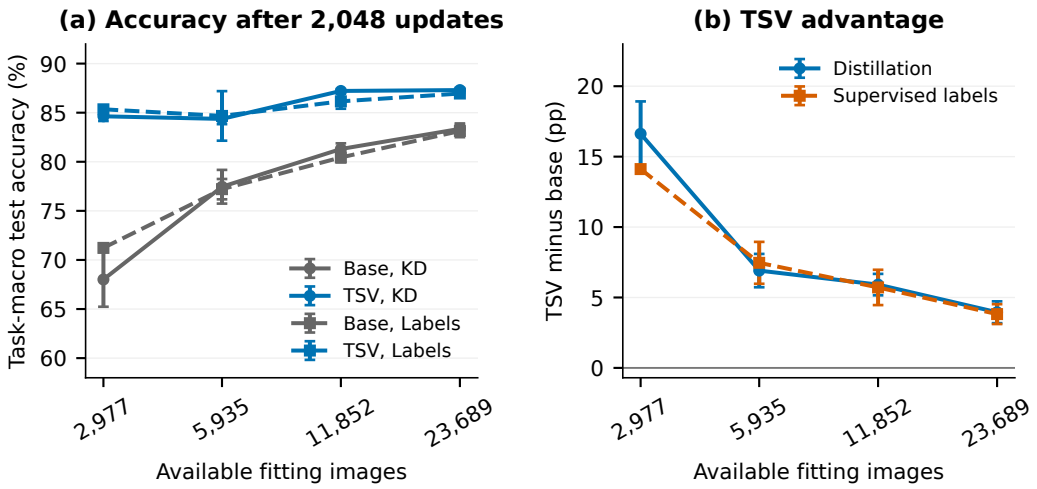

Figure 4: **A merged start helps under both training objectives:** ViT-B/32 at 2,048 updates, using a separate two-seed control campaign. (a) Task-macro accuracy from the pretrained base and TSV under distillation (KD) and supervised cross-entropy. (b) The initialization gap for each objective. Error bars are sample SDs across the two seeds, not confidence intervals. Every pool retains all 758 classes. These control means are not pooled with the five-seed estimates in Table 1.

The contrast is smaller and uncertain on ViT-B/16: On ViT-B/16, the same interventions reduce the gap by 2.92 ± 0.25 points for updates and 5.25 ± 2.46 points for data. Their paired difference is 2.33 points, with 95% interval $[-0.66, 5.32]$ and $p = 0.096$. Four of five seed-level contrasts favor data. The point estimate has the same direction as on ViT-B/32, but *this experiment does not resolve which intervention has the larger effect*.

Both initial gaps are smaller on ViT-B/16. With the small pool, its base reaches 77.54 rather than 70.39 on ViT-B/32; at the full-pool, 128-update setting, it reaches 83.79 rather than 80.32. This is consistent with the stronger base leaving less room for a merged initialization to help. It is not an isolated causal effect of backbone strength: changing the backbone also changes the specialists, representations, and cost per update. The cross-backbone difference in Δ is 4.62 points (Welch 95% interval $[0.14, 9.11]$, $p = 0.045$). This exploratory contrast between two backbones from one model family does not establish a scaling law.

A remaining advantage, not a measured closure point: At the largest tested data and update budgets, TSV still leads by 3.67 points on ViT-B/32 and 2.55 on ViT-B/16. We do not extrapolate a pool size at which the gap must disappear. The observed data response is nonlinear, the low-data points are seed-sensitive, and no measured setting closes the gap. Ratios of the two gap reductions describe these chosen interventions, not how many images a merge is intrinsically worth.

5 Controls and Interpretation

5.1 The advantage also appears with supervised training

The TSV initialization retains a similar gap under supervised and distilled adaptation at the full fitting-pool size. With 23,689 images, supervised training from the base reaches 83.13%, compared with 83.35% under distillation; supervised training from TSV reaches 86.95%, leaving a 3.82-point gap, versus 3.95 points under distillation. To test whether the primary result depends on teacher predictions in the loss, we repeat the ViT-B/32 data sweep using joint multi-task supervised fine-tuning with the same examples, updates, and learning rate. Figure 4 reports this separate two-seed comparison.

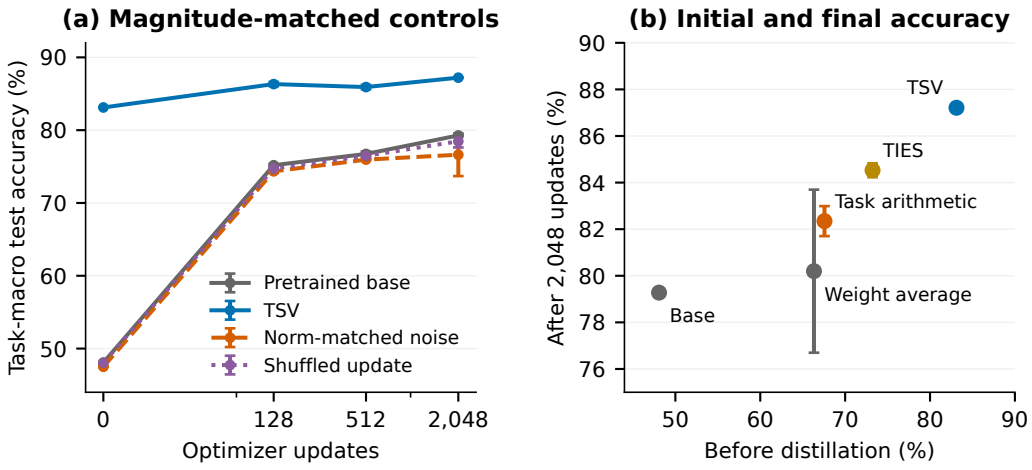

Figure 5: **The merged update carries more than its magnitude:** ViT-B/32, 6,064 fitting images, and two optimization seeds. (a) Random and shuffled norm-matched updates do not reproduce TSV's advantage. (b) Across five initializations, the ordering of mean accuracies is preserved after 2,048 updates. Error bars show sample SD; the large variation for averaging cautions against an inferential reading of the ordering. All accuracies use the full reused test collection.

This is not evidence that the objectives are interchangeable. At the smallest pool, supervised base training is 3.22 points better than distillation, and two seeds are insufficient for an equivalence claim. Nor is the supervised control a comparison with unrestricted joint training on all data used to produce the specialists. The specialists encode substantially more training exposure than the fitting pools available to either adapted model.

5.2 Update magnitude does not explain the benefit

Norm-matched Gaussian noise and shuffled update entries do not reproduce TSV's adaptation gain. Let $\delta = \theta_{\text{TSV}} - \theta_0$. For each tensor, we construct (i) isotropic Gaussian noise rescaled to the corresponding norm of δ, and (ii) a random permutation of that tensor's update entries. The latter preserves the entry distribution as well as the norm, while disrupting where the entries are placed. The recorded check verifies norm matching over all 199 floating-point tensors, to a relative tolerance of 10^{-3}. These are synthetic perturbations, not function-preserving permutations of whole neurons.

Both controls start near the base model's accuracy, not TSV's (Figure 5a). After adaptation, the noise and shuffled controls reach 76.63% and 78.47%, compared with 79.28% for the base and 87.21% for TSV. They therefore fail to reproduce the merged model's benefit. Thus preserving tensor norms, even together with marginal entry distributions, is insufficient to recover the benefit for these perturbations. It does not identify which features carry the useful information or show that merging is the only way to acquire it. The two controls each use one fixed perturbation draw; the reported seeds vary adaptation, not the synthetic initialization itself.

5.3 Optimization sensitivity and initialization quality

Learning rate and training loss: The tested learning-rate sweep does not eliminate the initialization gap or establish an optimal recipe. On the 6,064-image fitting set, a separate two-seed sweep compares learning rates 10^{-5}, 2×10^{-5}, and 5×10^{-5}. At 2,048 updates, both initializations attain their highest mean test accuracy among these rates at 10^{-5}, with a 7.94-point gap. This is a post hoc sensitivity check on the reused test set, not an independent hyperparameter-selection procedure.

At the primary ViT-B/32 full-pool endpoint, the mean terminal fitting loss is 0.0159 for the base and 0.0097 for TSV. A small loss does not establish convergence, and the base's loss is higher. Moreover, the ViT-B/32 reports attach the terminal loss to earlier saved checkpoints, so those fields

cannot support an intermediate-budget convergence claim. The accuracy response to additional updates is measurable; whether a different optimizer, schedule, or longer run would close the gap remains open.

Better starts remain better on average: With the same 6,064-image recipe, the mean accuracy ordering of the base, weight averaging, task arithmetic, TIES, and TSV is preserved after 2,048 updates (Figure 5b). The corresponding descriptive linear slope is 0.23: adaptation compresses, but does not remove, the initial spread. There are only five initializations and two seeds, so this is not a population-level correlation claim.

6 Discussion and limitations

What the results support: In this finite-budget setting, TSV provides an initialization that uses available fitting data more effectively than the pretrained base. On ViT-B/32, expanding the fitting pool reduces that advantage more than extending the full-pool update budget. ViT-B/16 shows a smaller point estimate and wider uncertainty relative to the effect. A useful practical comparison therefore reports both starting points across data and update budgets, rather than crediting a merge solely for beating a lightly trained base.

What the resource comparison does not price: The specialists are treated as available assets. Their original training is neither free nor charged to the adaptation budget. Our experiments do not compare merging with joint training under the same total data exposure or total compute. They also do not assign a cost to acquiring labels, caching teacher outputs, or selecting hyperparameters. A data-efficiency benefit in this access regime should not be read as an end-to-end efficiency guarantee. Nor do the measured budgets identify an equal-accuracy exchange between additional fitting images and optimizer updates.

Limits to generalization: The primary resource study uses one merging operator, two CLIP backbones, one expert per task and backbone, and eight image-classification tasks. Both expert sets come from the same publisher. The controls do not fully cross objective, merging method, and backbone, and several use only two seeds. Five-seed intervals remain sensitive to distributional assumptions. For example, an exact two-sided sign test of the five positive ViT-B/32 contrasts gives $p = 0.0625$, so not every test resolves the same significance threshold. All inferential comparisons are exploratory and unadjusted for multiple analyses. Reuse of the test collection is the strongest evaluation limitation. New expert sets, additional model families, and a prospectively frozen evaluation would be needed to establish broader resource laws.

Conclusion

A merged checkpoint can provide a useful data-efficient starting point for further adaptation, even when both initializations receive the same training signal. Its advantage is not explained by update magnitude alone, and its response to additional data and updates depends on the backbone and budget. The appropriate unit of value is therefore a measured gain under a stated access regime, not an unconditional claim that merging replaces training.

AI use statement

In this work, we used generative AI for manuscript organization and editing, and consistency checks of existing experimental results, literature lookup, and figure preparation and captions. We have reviewed all AI-assisted work to the best of our knowledge. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This study reuses public image-classification benchmarks and model checkpoints. It introduces no new collection of personal data. The narrow accuracy and retention measurements do not establish

fairness, safety, or suitability for deployment. Reuse of the target test collection and the limits of the resulting evidence are stated explicitly.

Reproducibility statement

We provide the code base used for experiments in this paper in teh supplementary zip file. Section 3 defines the objective, resource interventions, data roles, and uncertainty calculation. Appendix A specifies model and data details; Appendices B and C give seed-level results and controls. The accompanying manuscript source contains a numerical report snapshot and a script that regenerates the data figures, tables, and primary statistical contrasts without model inference.

References

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.04836.

Jun Chen, Qin Zhang, Weizhi Zhang, Xiao Luo, Philip Yu, and Ziyue Qiao. Learn to merge: Meta-learning for adaptive multi-task model merging. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=c11G7lmfhm.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017. doi: 10.1109/JPROC.2017.2675998. URL https://arxiv.org/abs/1703.00121.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, S. Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://www.robots.ox.ac.uk/~vgg/data/dtd/.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.00081.

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*, 2019. URL https://arxiv.org/abs/1709.00029.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

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. URL https://arxiv.org/abs/2106.09685.

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. URL https://arxiv.org/abs/2212.04089.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.08403.

Edan Kinderman, Itay Hubara, Haggai Maron, and Daniel Soudry. Foldable SuperNets: Scalable merging of transformers with different initializations and tasks. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2410.01483.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *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. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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. doi: 10.1109/5.726791.

Longhua Li, Lei Qi, Qi Tian, and Xin Geng. Model merging in the essential subspace. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31145–31154, 2026. URL https://arxiv.org/abs/2602.20208.

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

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2008.11687.

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. doi: 10.1109/IJCNN.2011.6033395.

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. URL https://arxiv.org/abs/2410.19735.

Mitchell Wortsman, Gabriel Ilharco, Samir Y. 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 *International Conference on Machine Learning*, pp. 23965–23998, 2022a. URL https://proceedings.mlr.press/v162/wortsman22a.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022b. URL https://arxiv.org/abs/2109.01903.

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

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.01708.

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. URL https://arxiv.org/abs/2310.02575.

A ADDITIONAL EXPERIMENTAL DETAILS

Task populations and prediction heads: Table 2 gives the pinned provider test populations. These counts describe the actual evaluation records, not necessarily every split distributed by each dataset's original authors. Accuracy is computed within each task, then averaged across tasks regardless of their population sizes. Each task uses fixed text-derived classification heads and its own class vocabulary. The model does not infer which task an image belongs to.

Table 2: Task populations in the pinned evaluation.

Task	Classes	Test images
Cars (Krause et al., 2013)	196	8,041
DTD (Cimpoi et al., 2014)	47	1,880
EuroSAT (Helber et al., 2019)	10	2,700
GTSRB (Stallkamp et al., 2011)	43	12,630
MNIST (LeCun et al., 1998)	10	10,000
RESISC45 (Cheng et al., 2017)	45	6,300
SUN397 (Xiao et al., 2010)	397	19,850
SVHN (Netzer et al., 2011)	10	26,032
Total	758	87,433

Checkpoints and preprocessing: The public pretrained checkpoints are:

ViT-B/32	`openai/clip-vit-base-patch32`
ViT-B/16	`openai/clip-vit-base-patch16`

Specialists are the `tanganke` releases for the matching backbone, with an underscore and task suffix appended to the model name. Cars uses the suffix `stanford-cars`; the others use the lowercase task names. Both backbones use the same cached 224-pixel CLIP image preprocessing and the same prompt templates. Text heads are constructed separately from each backbone's pretrained text tower. The pinned TSV implementation packs fused attention QKV before its matrix operation, averages nonmatrix task deltas, and uses scale 1 for the primary TSV start. This implementation convention is held fixed across the two backbones.

Fitting-pool construction: The 23,689-image pool is a superset of the original eight-images-per-class fitting role. It has up to 32 images per class, with fewer available images for some Cars classes. Class-stratified fractions retain all classes and give the four sizes in Table 6. The same subset construction is used for both initializations. Optimization seeds do not resample the fitting pool. An exploratory row-order subsampling scheme that removed whole classes was discarded and contributes no result to this paper.

Disjoint roles and exposure: The original fitting role has 6,064 images. The selection role and each of the two development roles have 3,032 images, or four per class. The extended fitting pool excludes every selection and development example. These identities are disjoint within the adaptation study, but the public specialists may have encountered all of TRAIN. The target test collection was already exposed to the project, including an earlier go/no-go decision about repair. Disjoint TRAIN development roles do not turn this reused test evaluation into fresh confirmation.

Optimization and resource accounting: The student has 87,456,000 trainable vision parameters in the full-weight ViT-B/32 runs. AdamW uses the settings in Section 3, with no learning-rate decay. A 128-update run has 32,768 student-image presentations; a 2,048-update run has 524,288. Presentations can repeat an image. Teacher logits are cached before student training. ViT-B/16 runs may split a 16-image task batch into smaller microbatches, weighted to preserve its mean gradient; the effective task and update weighting is unchanged. These presentation counts exclude teacher inference and are not FLOP measurements.

Separate experimental campaigns: The primary ViT-B/32 comparison uses the five-seed anchored campaign, not the earlier two-seed data sweep. The ViT-B/16 comparison pools seeds 117–

Table 3: Every paired primary contrast, in percentage points. C is gap reduction from more updates, D is gap reduction from more fitting images, and $\Delta = D - C$. All five seeds are included.

	ViT-B/32			ViT-B/16		
Seed	C	D	Δ	C	D	Δ
117	4.32	11.59	7.28	3.02	6.04	3.02
118	2.46	13.75	11.29	2.62	1.31	-1.31
119	3.19	6.49	3.30	3.17	5.19	2.01
120	2.92	12.25	9.33	2.67	8.05	5.38
121	3.49	7.08	3.59	3.10	5.66	2.57

Table 4: The complete ViT-B/32 two-pool, three-budget grid. Entries are mean ± sample SD over the same five seeds. Accuracy is measured on the full reused test collection. The low-data base improves at 512 updates but declines by 2,048, so longer training is not uniformly beneficial. Each accuracy column is ranked within its fitting-pool block.

Fit images	Updates	Base (%)	TSV (%)	Gap (pp)
2,977	128	69.94 ± 0.61	85.62 ± 0.26	15.68 ± 0.60
2,977	512	72.55 ± 0.34	86.32 ± 0.19	13.77 ± 0.33
2,977	2,048	70.39 ± 3.32	84.30 ± 0.47	13.91 ± 3.58
23,689	128	80.32 ± 0.48	87.27 ± 0.15	6.95 ± 0.49
23,689	512	82.77 ± 0.24	87.49 ± 0.18	4.72 ± 0.24
23,689	2,048	83.64 ± 0.40	87.32 ± 0.12	3.67 ± 0.47

118 from the data and update campaigns with seeds 119–121 from the extension. Duplicate full-pool endpoints are checked for agreement and counted once. The supervised data sweep, magnitude controls, and initializer-quality experiment use seeds 117–118. The parameter-efficient repair study uses its own three-seed protocol. These campaigns should not be combined merely because some seed identifiers or nominal settings coincide.

B PRIMARY RESULTS AND UNCERTAINTY

Figure 3 in the main text reports both full-pool training curves. The tables below provide the underlying seed-level and per-task detail.

The estimand: Equations 5 and 6 use the same terminal gap. Hence

$$\Delta_s = G_s(2{,}977, 2{,}048) - G_s(23{,}689, 128).$$

This cancellation is useful for checking the calculation, but it also makes the scope clear: the contrast compares two selected resource settings. It is not an economic exchange rate or a continuous data-by-compute response law. Table 1 reports both starting gaps and the shared endpoint so that the reduction is not mistaken for absolute accuracy.

A common-start check on ViT-B/32: Table 4 reports every point in the two-pool, three-budget grid. Let $G_s^0 = G_s(2{,}977, 128)$. The gap reductions from a common low-resource start are

$$C_s^0 = G_s^0 - G_s(2{,}977, 2{,}048), \qquad D_s^0 = G_s^0 - G_s(23{,}689, 128).$$

Their difference satisfies $D_s^0 - C_s^0 = \Delta_s$ exactly. In this parameterization, more updates close 1.77 ± 3.02 points and more data close 8.73 ± 0.70 points. Thus the separate reductions depend on the reference point, but their paired contrast does not. This identity does not supply an additional significance test or make the two resource increases cost-equivalent.

Seed-level inference: For each backbone, we form Δ_s within seed and report

$$\bar{\Delta} \pm t_{0.975,4} \frac{\mathrm{SD}(\Delta_s)}{\sqrt{5}}.$$

Table 5: Per-task accuracy at the primary full-pool, 2,048-update endpoint. Entries are means across five seeds. Each task column ranks the two initializations within a backbone, never different tasks against each other. Gap rows are unshaded differences of unrounded means, not per-task significance tests.

Initialization	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC45	SUN397	SVHN
ViT-B/32								
Base	72.38	72.15	94.21	92.85	95.91	85.95	70.25	85.44
TSV	73.66	74.93	96.93	97.36	99.23	91.32	70.42	94.68
Gap (pp)	+1.28	+2.78	+2.72	+4.51	+3.32	+5.37	+0.17	+9.24
ViT-B/16								
Base	80.80	75.36	94.52	94.58	98.12	89.29	74.59	91.74
TSV	84.49	77.74	97.34	97.45	99.25	92.87	74.75	95.51
Gap (pp)	+3.69	+2.38	+2.82	+2.87	+1.12	+3.57	+0.16	+3.76

Table 6: ViT-B/32 objective comparison at 2,048 updates. Each entry is mean $\pm$ sample SD over two seeds in the supporting data-sweep campaign. All values are task-macro accuracies (%) on the reused test collection. Rankings compare fitting-pool sizes within each objective/initialization column.

	Distillation		Supervised labels	
Fit images	Base	TSV	Base	TSV
2,977	68.01 ± 2.78	84.63 ± 0.48	71.23 ± 0.02	85.35 ± 0.17
5,935	77.46 ± 1.72	84.37 ± 0.53	77.21 ± 1.04	84.67 ± 2.52
11,852	81.30 ± 0.58	87.21 ± 0.18	80.44 ± 0.50	86.15 ± 0.75
23,689	83.35 ± 0.55	87.30 ± 0.22	83.13 ± 0.62	86.95 ± 0.08

The corresponding two-sided one-sample tests give $p = 0.011$ for ViT-B/32 and $p = 0.096$ for ViT-B/16. We first form the initialization gap within each matched condition and then form resource contrasts within seed. The two starts share a batch sequence at a given pool size; pairing does not imply identical sampled images across different pool sizes. With only five seeds, the approximate sampling assumptions matter. Exact two-sided sign tests give $p = 0.0625$ for the five positive ViT-B/32 contrasts and $p = 0.375$ for the four positive ViT-B/16 contrasts. The evidence should therefore be read through the effect sizes and individual seeds, not as a claim that every reasonable test establishes significance.

Small-sample and cross-backbone sensitivity: A percentile bootstrap of the five ViT-B/16 seed differences gives $[0.24, 4.15]$, whereas its $t(4)$ interval is $[-0.66, 5.32]$. We use the same t procedure for both backbones rather than select the interval that excludes zero. Neither estimator removes the uncertainty caused by five observations. The difference between backbone means is 4.62 points. The sealed comparison reports a two-sided Welch $p = 0.045$; its approximate 95% interval is $[0.14, 9.11]$ points. This is an exploratory cross-backbone contrast, not a general law of model scale. No reported test is adjusted for the number of exploratory analyses.

The gap is not uniform across tasks. At the ViT-B/32 full-pool endpoint it is 9.24 points on SVHN but only 0.17 on SUN397; the corresponding ViT-B/16 gaps are 3.76 and 0.16. Equal task weighting is therefore consequential. These per-task results do not support treating all 87,433 images as independent replications of the central resource claim.

C Supporting controls

Objective comparison: Only the adaptation loss changes between the distillation and supervised campaigns. The supervised model uses the same frozen task heads, rather than learning a new classifier for each task. At full data, the matched supervised gap is 3.82 points; comparing supervised base training with distilled TSV instead would mix objectives and answer a different question. Table 6 retains the two-seed campaign values rather than substituting five-seed endpoints from the primary study.

Table 7: Learning-rate sensitivity on ViT-B/32 with 6,064 fitting images. Each value is mean reused-test accuracy (%) over two seeds. The grid was inspected using these test outcomes and is not an independent selection experiment. Each column ranks rates at a fixed update budget and initialization.

	128 updates		512 updates		2,048 updates	
Learning rate	Base	TSV	Base	TSV	Base	TSV
1e-05	75.19	86.33	76.74	85.92	79.28	87.21
2e-05	74.48	83.93	73.90	82.45	73.17	81.51
5e-05	67.74	76.33	64.03	71.27	60.82	64.42

Table 8: Magnitude-matched initialization controls. ViT-B/32, 6,064 fitting images, and the same distillation recipe. Initial accuracies are deterministic checkpoint measurements; final accuracies are means $\pm$ sample SD over two optimization seeds on the full reused test collection. Each accuracy column ranks all four matched initializations, including the pretrained control.

Initialization	Before training (%)	After 2,048 updates (%)
Pretrained base	48.08	79.28 ± 0.16
Norm-matched noise	47.50	76.63 ± 2.94
Shuffled update	48.04	78.47 ± 0.85
TSV	83.12	87.21 ± 0.04

Loss-recording scope: In the ViT-B/32 update runner, each saved checkpoint's report row carries the mean loss from the *final* 50 updates of the complete run. That field is therefore a terminal loss, even when the row identifies an earlier checkpoint. We use it only at 2,048 updates. In the five-seed anchored campaign, the terminal base/TSV losses are 0.0065/0.0005 at the small pool and 0.0159/0.0097 at the full pool. Loss values across different fitting pools or objectives are not convergence certificates or directly comparable generalization measures. The earlier-checkpoint accuracy measurements remain usable independently of this logging convention.

Synthetic updates: If tensor j has update δ_j, the noise control uses $\tilde{\delta}_j = \|\delta_j\|_F z_j / \|z_j\|_F$ for an isotropic Gaussian z_j. The shuffled control randomly reorders the entries of δ_j. Both are added to the pretrained base. The recorded total update norm is 5.9002, and the numerical matching check covers 199 floating-point tensors. The perturbation generator uses seed 20461; optimization then uses seeds 117 and 118. The experiment tests these two draws, not an average over a population of random initializations.

Initialization quality: The quality sweep evaluates the pretrained base and four fixed merge checkpoints. At 2,048 updates, their mean accuracies are 79.28, 80.20, 82.35, 84.53, and 87.21% for the base, averaging, task arithmetic, TIES, and TSV, respectively. The Spearman correlation of these five mean endpoints with pre-adaptation accuracy is 1.00, and the fitted linear slope is 0.23. These are descriptive summaries of the chosen initializations, not evidence of a universal predictive relation or a guarantee that each seed has the same ordering.

D BENCHMARK AND PARAMETER-EFFICIENT REPAIR

Scope of the benchmark: Table 9 reports the existing all-eight-task ViT-B/32 benchmark. Each merge is a single vision checkpoint. Specialist and score ensemble rows are references, not comparable single-checkpoint methods. Specialist-normalized accuracy divides each task's accuracy by its own specialist accuracy before taking the task mean; it is not an oracle score. The score ensemble's 77.14% illustrates that a naive ensemble is not an upper bound.

The AdaMerging and local AdaMerging++ implementations change the published data-access regime by fitting coefficients on TRAIN calibration data. They are not faithful reproductions of transductive AdaMerging (Yang et al., 2024). The benchmark does not cover all contemporary

Table 9: Pinned ViT-B/32 benchmark on the full reused test collection. AdaMerging variants use TRAIN calibration, not the published transductive TEST procedure. Their error bars summarize three coefficient-optimization seeds. "Norm." denotes specialist-normalized task-macro accuracy. Rankings exclude the specialist and ensemble reference rows.

Method	Accuracy (%)	Norm. (%)
Weight average	66.33	73.71
Task arithmetic	67.57	74.28
TIES	73.23	80.66
AdaMerging (TRAIN-cal.)	80.29 ± 0.09	88.67
AdaMerging++ (TRAIN-cal.)	80.44 ± 0.03	88.68
TSV	83.12	91.57
Task specialists (reference)	90.38	100.00
Score ensemble (reference)	77.14	n/a

Table 10: TSV repair on the full reused test collection. Adaptation uses 6,064 fitting images and 128 updates; error bars are sample SD over three seeds. Low-rank arms receive a two-rate search, whereas the full-weight reference uses one previously fixed rate. This is not a search-matched comparison. Only the three parameter-matched low-rank arms are ranked; equal displayed accuracies share a color.

TSV-initialized model	Trainable parameters	Accuracy (%)
TSV, no repair	0	83.12
Full-weight KD	87,456,000	86.44 ± 0.28
Retained spectral coordinates	589,824	87.17 ± 0.10
Random subspace	589,824	87.31 ± 0.02
Native LoRA	589,824	87.31 ± 0.14

merging methods, including ESM and MetaMerging (Li et al., 2026; Chen et al., 2026). TSV's lead within this table is not a state-of-the-art claim.

Repair parameterizations: All repair arms start from the same TSV checkpoint. For each eligible attention or MLP matrix $W \in \mathbb{R}^{m \times n}$, the spectral arm learns

$$\Delta W = UBAV, \qquad U^\top U = I_k, \quad VV^\top = I_k, \quad B \in \mathbb{R}^{k \times r}, \quad A \in \mathbb{R}^{r \times k}, \quad r = 8.$$

The bases U, V are fixed retained TSV coordinates. The random-subspace arm uses fixed random orthonormal bases of the same dimensions. Native LoRA (Hu et al., 2022) learns an unrestricted low-rank product in the original matrix coordinates, with per-matrix ranks chosen to match the factor budget. Each arm has 589,824 trainable factors. Fused QKV is treated as one matrix; embeddings and nonmatrix parameters are frozen in the low-rank arms.

Selection and results: Low-rank arms use learning rates 10^{-4} and 10^{-3}, selected by mean accuracy over all three seeds on the TRAIN selection role. Spectral and native LoRA select 10^{-3}; the random subspace selects 10^{-4}. Full-weight KD uses the previously fixed 10^{-5}. A separate TRAIN check preceded the frozen reused-test evaluation, without selecting a best seed.

Native LoRA exceeds spectral repair by 0.14 points, with a conditional paired image-bootstrap interval of $[-0.14, 0.39]$. It differs from random-subspace repair by less than 0.01 points. These results provide no resolved advantage for the retained spectral coordinates, rather than proving exact equivalence. Native LoRA exceeds full-weight KD by 0.87 points, with interval $[0.53, 1.24]$, but the learning-rate search and trainable parameter scope differ. All image-bootstrap intervals fix the tasks and fitted checkpoints; they are not the five-seed t intervals used for the primary resource study.

Parameter counts are not deployment savings: The low-rank factor count is 0.674% of the full-weight ViT-B/32 count. However, each spectral or random frame set stores 113,246,208 frozen scalars, approximately 453 MB in float32. Selected-rate training times in the recorded environment average 173, 172, and 157 seconds for spectral, random, and native LoRA, respectively. Every up-

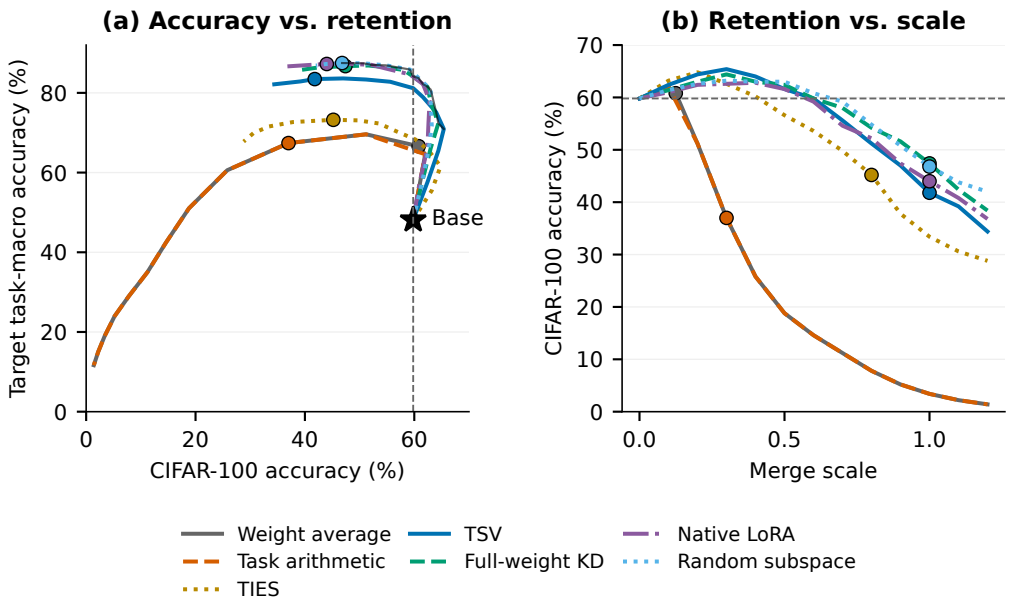

Figure 6: An exploratory interpolation sweep. (a) Target accuracy against CIFAR-100 zero-shot accuracy. Circles mark the benchmark reference scales, the star is the pretrained base, and the black line connects the empirical nondominated points. (b) Retention against interpolation scale. The dashed reference lines mark base retention. This extends the evaluation of an established interpolation baseline, not the claim that interpolation is new.

date is folded into a dense vision checkpoint of about 350 MB. The smaller trainable count therefore does not establish a smaller deployed model or an end-to-end speed advantage.

Why the controls are necessary: For semi-orthogonal U, V, $\|UCV\|_F = \|C\|_F$. If both bases are square, rotating a rank-r product into and out of these coordinates preserves exactly the ordinary rank-r update family. Rectangular bases instead restrict its row and column spaces. These are elementary coordinate facts, not new theorems. They motivate a comparison with random subspaces and native LoRA before attributing gains to a particular spectral basis.

E TARGET ACCURACY AND RETAINED ZERO-SHOT ACCURACY

For a merged or repaired checkpoint, we sweep the scale in $\theta(\alpha) = \theta_0 + \alpha(\theta_{\text{merged}} - \theta_0)$. The reference point is the checkpoint's selected operating scale in our benchmark, not necessarily a scale taken unchanged from a publication. We measure target accuracy on a fixed 1,000-image-per-task test subset and retention on 500 CIFAR-100 images (Krizhevsky, 2009) that had not been used by the project before this sweep. The pretrained base scores 48.14% on this target subset and 59.8% on the retention probe. The target values here are not the full-test values in Table 9.

At the TSV reference scale of 1.0, target accuracy is 83.5% and CIFAR-100 accuracy is 41.8%. At scale 0.6, they are 81.2% and 59.8%: an 18.0-point retention recovery for 2.3 points of target accuracy. Six of the seven reference operating points are dominated by another measured point on the empirical two-axis frontier. This is point-estimate dominance, not a statistically certified ordering.

Retention is nonmonotonic over the measured grid. TSV at scale 0.3 reaches 65.4% on CIFAR-100, above the base's 59.8%. Interpolation with the base is therefore an important control, consistent with the motivation of WiSE-FT (Wortsman et al., 2022b). The 500-image, single-dataset probe is too small and narrow to establish a general-capability improvement. Using it to describe the frontier also does not create an independent validation set for a scale chosen after observing that frontier.

Table 11: Reference scales and the largest measured scale that retains at least the base model's CIFAR-100 accuracy. All accuracies are percentages. "Target at recovery" is measured on the fixed target subset. Each accuracy column is ranked separately; interpolation scales are not ranked.

	Reference point			Recovery point	
Method	Scale	Target	CIFAR	Scale	Target
Weight average	0.125	66.6	60.8	0.125	66.6
Task arithmetic	0.3	67.4	37.0	0.1	64.6
TIES	0.8	73.2	45.2	0.4	68.3
TSV	1	83.5	41.8	0.6	81.2
Random subspace	1	87.5	46.8	0.7	85.9
Native LoRA	1	87.2	44.0	0.6	84.6
Full-weight KD	1	86.7	47.4	0.6	84.1

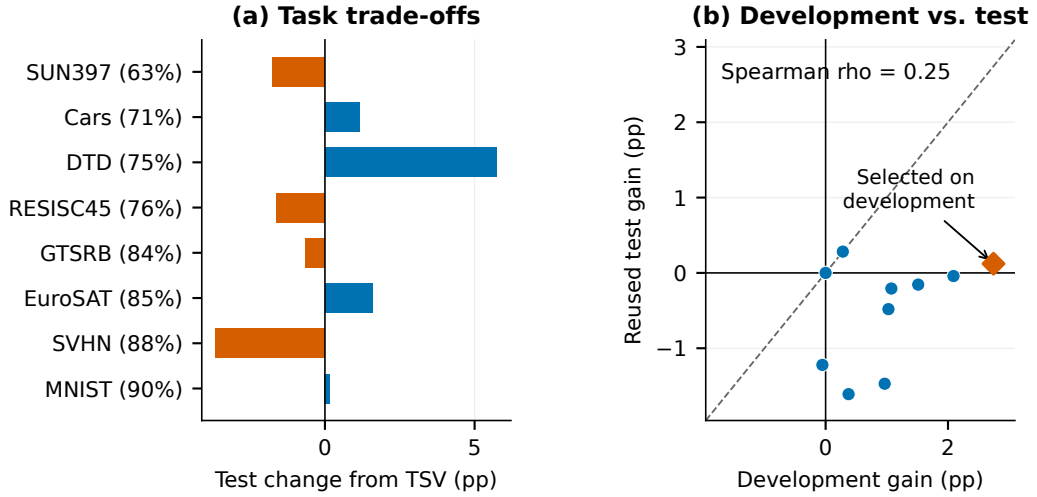

Figure 7: A fixed-budget spectral reallocation does not deliver a resolved aggregate test gain. (a) Per-task accuracy changes for the development-selected allocation; percentages beside task names give retained spectral energy under uniform TSV. (b) Development and reused-test gains for ten allocation arms. The diamond is the development-selected arm; the dashed diagonal denotes identical gains on both roles.

F Spectral allocation and development-set transfer

A fixed-budget allocation control: TSV's uniform rank allocation retains unequal fractions of task spectral energy, from 62.7% for SUN397 to 89.6% for MNIST. A data-free reallocation that equalizes retained energy redistributes the same total rank budget. The selected allocation improves task-macro accuracy by only 0.12 points on the full reused test collection, with a conditional paired interval of $[-0.09, 0.34]$. Gains on some tasks, including DTD, are offset by losses on others, including SVHN (Figure 7). Example-weighted accuracy decreases by about 1.4 points. This is an observed trade-off for the evaluated allocations, not a conservation law for merging.

Development agreement does not ensure test transfer: The same selected allocation improves accuracy by 2.74 and 1.85 points on two disjoint class-balanced TRAIN development roles before giving the unresolved 0.12-point test gain. Arm rankings have Spearman correlation 0.84 between the development roles but only 0.25 between development and test. Both development roles share the same four-examples-per-class construction, whereas the test population has different within-task class frequencies. Shared construction is a plausible contributor to the mismatch, not an isolated

causal explanation. Replication across similarly constructed development splits does not by itself validate transfer to the target test distribution.