ADA-HISS: SPECTRAL–SPARSE MODEL MERGING IN THE SMALL-ROOM REGIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging consolidates multiple fine-tuned experts into a single shared backbone without joint retraining. Most merging rules, however, are largely developed in dense, overparameterized regimes, where models have enough redundancy to absorb interference among task updates. We consider a deployment-oriented setting in which task experts are fine-tuned from a structurally pruned backbone rather than from a dense model, reflecting edge and resource-constrained scenarios. In this setting, we discover the *Small-Room* problem where pruning reduces the effective degrees of freedom available for consolidation, causing task updates to interfere more as the number of experts grows. To address this problem, we propose **Ada-HiSS**, an **Ada**ptive **H**ybr**i**d **S**pectral–**S**parse merging method. Ada-HiSS decomposes each task update into a globally allocated low-rank spectral core and a sparse residual tail. The spectral core captures dominant low-rank structure in each task update, while the sparse tail preserves localized task-specific changes discarded by the spectral core. We evaluate Ada-HiSS in two complementary settings: shared-backbone merging and task-update compression. In both settings, Ada-HiSS substantially improves over existing parameter-space and spectral baselines, with the largest gains on interference-sensitive fine-grained and scene-recognition tasks.

1 INTRODUCTION

The pretrain–fine-tune paradigm has produced large collections of task-specialized models. A common workflow is to start from a strong pretrained backbone and fine-tune separate experts for downstream datasets. This is convenient, but serving one checkpoint per task scales linearly in storage, memory bandwidth, and operational complexity. Multi-task training can avoid this duplication, but it typically requires joint access to all task data and substantial retraining compute (Caruana, 1997; Kendall et al., 2018). These constraints motivate *data-free model merging*: given a shared initialization and several fine-tuned experts, combine their parameter updates into a single deployable model without accessing the original training data.

Existing model-merging methods include weight averaging and model soups (Wortsman et al., 2022), task-vector arithmetic (Ilharco et al., 2023), interference-aware coordinate rules such as TIES (Yadav et al., 2023), tuning-free mask/rescale schemes such as EMR (Huang et al., 2024), and more structured consolidation methods based on localized task supports or channels (Wang et al., 2024; Zhang et al., 2025). Most of these studies evaluate dense, overparameterized backbones. In parallel, practical deployment increasingly relies on structured pruning, which removes channels, widths, or coupled parameter groups to obtain real memory and latency reductions (Liu et al., 2017; He et al., 2017; Yu et al., 2019; Fang et al., 2023). This deployment shift exposes a different merging regime where experts may be fine-tuned from a backbone whose redundant capacity has already been removed.

Our central observation is that pruning changes the geometry of model merging. Dense models can often absorb interference because many parameters are redundant or weakly load-bearing. Structured pruning removes much of that slack. The remaining parameterization is a "small room" in which multiple task updates must coexist. Coordinate-wise merging, even with sign conflict resolution, can fail because the same reduced set of parameters must encode more task-specific changes.

This paper proposes **Ada-HiSS**, a representation-level merging method designed for this capacity-constrained regime. Instead of merging task updates as flat coordinate vectors, Ada-HiSS decomposes

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.

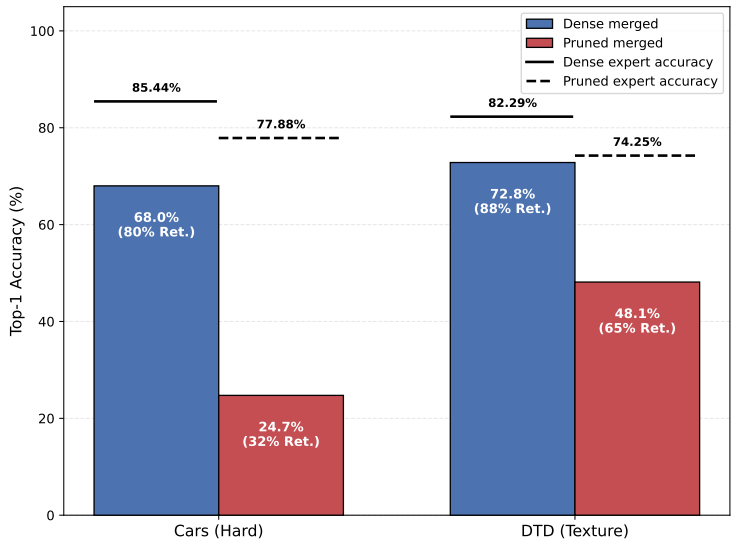

Figure 1: **Controlled Small-Room diagnostic.** Applying the same coordinate-wise TIES rule to dense and structured-pruned experts shows lower retention after pruning. The dashed lines show single-task expert accuracy, bars show merged accuracy, and text shows retention. This diagnostic isolates the pruning/merging interaction; the full multi-task results are reported in Tables 3–4.

each matrix-shaped update into two complementary pieces: a low-rank *spectral core* and a sparse *residual tail*. The spectral core is allocated globally across layers and tasks by a single singular-value threshold. This spends rank where the update spectrum says it is most useful, rather than imposing the same rank in every layer. The sparse tail then stitches back a small number of high-magnitude residual entries, preserving sharp task-specific details that low-rank truncation may omit.

Contributions. We make three contributions. First, we identify the *Small-Room* failure mode induced by structured pruning and formalize why cross-task interference grows when the effective update space shrinks. Second, we introduce Ada-HiSS as a spectral–sparse data-free merging framework where a globally allocated spectral core captures matrix-valued task structure, while a compact sparse residual tail preserves localized corrections. Third, we evaluate Ada-HiSS under two complementary protocols: the Shared-Backbone Protocol studies deployment-oriented merging into one shared backbone plus compact task-specific tails across dense and pruned ViT-B/16 model families, while the Task-Update Compression Protocol isolates how well the spectral–sparse representation stores each task update under a compact budget.

2 Related Work

Model merging. Model soups average the weights of fine-tuned models to improve accuracy without increasing inference cost (Wortsman et al., 2022). Task arithmetic represents fine-tuning as an additive task vector and composes edits in weight space (Ilharco et al., 2023). TIES-Merging identifies two important coordinate-level interference sources—redundant updates and sign disagreements—and resolves them through trimming, sign election, and aligned merging (Yadav et al., 2023). EMR-Merging elects a unified model and uses lightweight task-specific masks/rescalers, while remaining tuning-free and data-free (Huang et al., 2024). Other work studies Fisher-weighted merging (Matena & Raffel, 2022), model re-basin and permutation symmetries (Ainsworth et al., 2023), localized task information (Wang et al., 2024), and channel-level specialization in merged experts (Zhang et al., 2025).

Spectral structure of task updates. Task singular-vector methods show that task updates often contain useful low-rank layer-level structure, and that treating task vectors as matrices rather than flat coordinate vectors can reduce interference (Gargiulo et al., 2025). Ada-HiSS builds on this

spectral perspective, but introduces two concrete changes. First, instead of assigning a fixed rank independently to each layer, Ada-HiSS allocates spectral components globally across tasks and layers, allowing rank to follow the singular-value distribution of the full merge problem. Second, Ada-HiSS augments the spectral core with a sparse residual tail, preserving high-magnitude localized edits that are not represented by the retained low-rank components.

Structured pruning. Structured pruning removes entire channels, groups, or coupled structures to produce dense executable subnetworks with real hardware benefits (Liu et al., 2017; He et al., 2017; Yu et al., 2019). DepGraph provides an automatic dependency graph for generic structural pruning across architectures, including transformers (Fang et al., 2023). Isomorphic pruning further studies importance comparison across heterogeneous structures in modern vision models (Fang et al., 2024). Our work is complementary to pruning algorithms: we study the model-merging problem that arises after multiple task experts have been fine-tuned from the same structurally pruned backbone.

3 WHY THE PRUNED REGIME FAILS: THE SMALL-ROOM EFFECT

Let θ_0 be the shared pruned backbone and θ_t the expert fine-tuned for task $t \in \{1, \dots, T\}$. Define the task update

$$\Delta_t = \theta_t - \theta_0. \tag{1}$$

A generic additive merge constructs

$$\theta_{\text{merge}} = \theta_0 + \delta, \qquad \delta = A(\Delta_1, \dots, \Delta_T), \tag{2}$$

where A is a data-free aggregation operator.

3.1 A LOCAL INTERFERENCE MODEL

To expose the failure mode, approximate task t around its fine-tuned expert by a local quadratic:

$$\mathcal{L}_t(\theta_0 + \delta) \approx \mathcal{L}_t(\theta_t) + \frac{1}{2}\|\delta - \Delta_t\|^2_{H_t}, \tag{3}$$

where $H_t \succeq 0$ is a local curvature matrix and $\|x\|^2_{H_t} = x^\top H_t x$. To isolate the interference term, consider a simple scaled sum $\delta = \alpha \sum_{j=1}^{T} \Delta_j$. The excess quadratic error for task t is

$$\|\delta - \Delta_t\|^2_{H_t} = \left\| (\alpha - 1)\Delta_t + \alpha \sum_{j \neq t} \Delta_j \right\|^2_{H_t}. \tag{4}$$

The second term is the interference from all other tasks. Coordinate-wise methods such as trimming or sign resolution attempt to reduce this term in the original parameter coordinates. The Small-Room problem is that, after pruning, there are fewer effective directions in which these task updates can avoid one another. Eq. (4) isolates the effect studied by the Small-Room model: pruning reduces the effective dimension in which cross-task updates can be arranged, so the projection of other task updates onto task t's direction grows as more experts are merged.

Proposition 1 (Projected interference under an isotropic Small-Room model)**.** *Let $\mathcal{U}$ be an r-dimensional effective update subspace, and let $z_1, \dots, z_T$ be independent unit vectors sampled isotropically from the unit sphere in $\mathcal{U}$. For task t, define the squared projected interference from the other tasks as*

$$I_t = \left\langle z_t, \sum_{j \neq t} z_j \right\rangle^2. \tag{5}$$

Then

$$\mathbb{E}[I_t] = \frac{T - 1}{r}. \tag{6}$$

Consequently, under the same isotropic normalized-update model, reducing the effective dimension from r_{dense} to r_{pruned} increases this expected squared projection by a factor $r_{\text{dense}}/r_{\text{pruned}}$.

The result shows that as the effective update space becomes smaller, random or weakly aligned task updates have larger expected overlap, leaving coordinate-wise merging rules with less slack to resolve collisions. The proof is given in Appendix A.

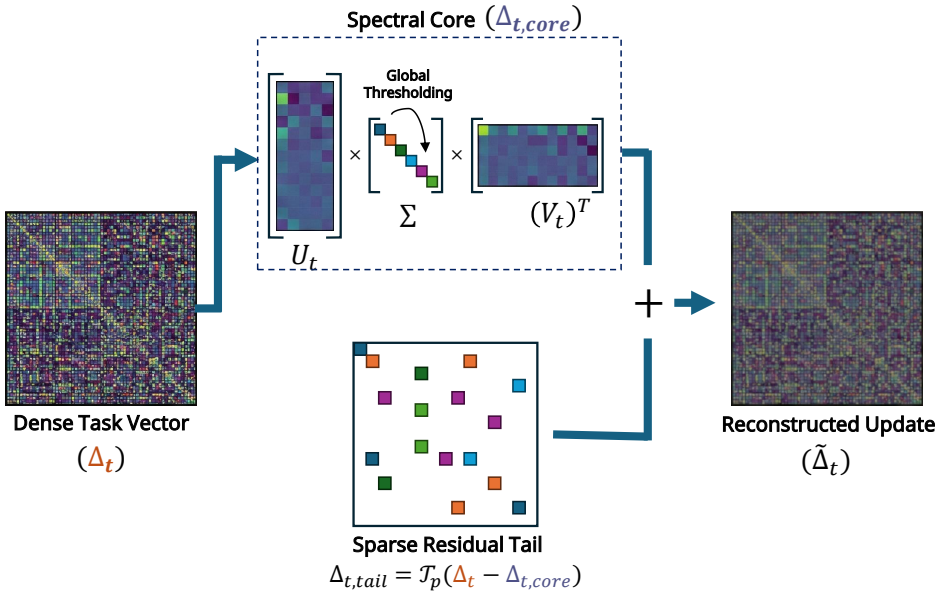

Figure 2: **Ada-HiSS schematic.** Ada-HiSS decomposes each matrix-shaped task update Δ_t into two complementary components: a globally thresholded low-rank spectral core $\Delta_{t,\text{core}}$ and a sparse residual tail $\Delta_{t,\text{tail}} = \mathcal{T}_p(R_t)$, where R_t is the protocol-specific residual. In the shared-backbone protocol, aligned spectral cores form the shared backbone while the sparse tail is stored as task-specific state and applied only for the corresponding task. In the task-update compression protocol, $\Delta_{t,\text{core}} + \Delta_{t,\text{tail}}$ is stored per task and applied to the base model only for that task.

4 Ada-HiSS: Adaptive Hybrid Spectral–Sparse Merging

Ada-HiSS targets the failure in Eq. (4) by changing the representation of each task update before composition. The method has four steps: compute matrix-level spectral decompositions of the task updates, allocate a global spectral budget, retain a compact sparse residual tail, and compose the resulting spectral–sparse representation according to the evaluation protocol. See Fig. 2.

4.1 Problem Setup

For a matrix-shaped parameter tensor in layer ℓ and task t, write the task update as $\Delta_t^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$. In ViT-style architectures, ℓ indexes the corresponding matrix-shaped projection tensors, such as attention and MLP projections. Vector-shaped parameters such as biases and normalization parameters are handled separately with a simple scaled sum. For matrices, compute the compact SVD:

$$\Delta_t^{(\ell)} = U_t^{(\ell)} \Sigma_t^{(\ell)} (V_t^{(\ell)})^\top = \sum_{i=1}^{q_\ell} \sigma_{t,i}^{(\ell)} u_{t,i}^{(\ell)} (v_{t,i}^{(\ell)})^\top, \quad q_\ell = \min(m_\ell, n_\ell). \tag{7}$$

4.2 Global Spectral Allocation

Rather than selecting the same rank for each layer, Ada-HiSS collects singular values across all matrix-shaped layers and all tasks:

$$\mathcal{S} = \left\{ \sigma_{t,i}^{(\ell)} : t = 1, \ldots, T,\ \ell \in \mathcal{M},\ i = 1, \ldots, q_\ell \right\}. \tag{8}$$

Given a spectral budget K (or a budget fraction k), choose a threshold τ so that the retained set contains the top K singular values in $\mathcal{S}$. The spectral core is

$$\Delta_{t,\text{core}}^{(\ell)} = \sum_{i:\, \sigma_{t,i}^{(\ell)} \geq \tau} \sigma_{t,i}^{(\ell)} u_{t,i}^{(\ell)} (v_{t,i}^{(\ell)})^\top. \tag{9}$$

In implementation, each matrix layer may retain at least one component to avoid degenerate zero cores.

Proposition 2 (Global thresholding is optimal for Frobenius SVD reconstruction)**.** *For each task-layer matrix A_m, where $m = (t, \ell)$, let*

$$A_m = \sum_i \sigma_{m,i} u_{m,i} v_{m,i}^\top \tag{10}$$

be its compact SVD. Consider approximations of the form

$$\widehat{A}_m = \sum_i x_{m,i} \sigma_{m,i} u_{m,i} v_{m,i}^\top, \qquad x_{m,i} \in \{0, 1\}, \tag{11}$$

with a global component budget $\sum_{m,i} x_{m,i} \le K$. The choice minimizing

$$\sum_m \|A_m - \widehat{A}_m\|_F^2 \tag{12}$$

is obtained by retaining the K largest singular values globally, with arbitrary tie-breaking.

This motivates using a global singular-value threshold rather than a fixed per-layer rank when allocating the spectral budget; see Appendix A for further details.

4.3 Sparse residual tail

The spectral core captures high-energy low-rank structure, but fine-grained tasks may depend on sharp, localized residual edits. Ada-HiSS therefore stores a sparse tail by applying hard thresholding to a protocol-specific residual. In the Shared-Backbone Protocol, after the shared spectral vector Δ_{shared} is formed as in Eq. (18), the task-specific residual is

$$R_t^{\text{SB}} = \Delta_t - \Delta_{\text{shared}}, \tag{13}$$

and Ada-HiSS stores only the top-magnitude residual coordinates,

$$\Delta_{t,\text{tail}}^{\text{SB}} = \mathcal{T}_p(R_t^{\text{SB}}). \tag{14}$$

In the Task-Update Compression Protocol, the residual is instead defined per matrix layer as

$$R_t^{(\ell)} = \Delta_t^{(\ell)} - \Delta_{t,\text{core}}^{(\ell)} \tag{15}$$

and the compressed task update keeps the top s_ℓ residual entries,

$$\Delta_{t,\text{tail}}^{(\ell)} = \mathcal{T}_{s_\ell}(R_t^{(\ell)}). \tag{16}$$

Here $\mathcal{T}_s(\cdot)$ zeros all but the s largest-magnitude entries. The sparse budget can be specified as a global fraction p of residual entries or as per-layer counts.

Proposition 3 (Hard thresholding is optimal for sparse residual reconstruction)**.** *Let $R \in \mathbb{R}^{m \times n}$ be a residual matrix and let s be a sparsity budget. A solution of*

$$\min_{\|Z\|_0 \le s} \|R - Z\|_F^2 \tag{17}$$

is obtained by keeping the s entries of R with largest absolute value and setting all other entries to zero, i.e. $Z = \mathcal{T}_s(R)$. Ties may be broken arbitrarily.

This motivates using top-magnitude residual entries as the sparse tail in both protocols; see Appendix A for the proof.

4.4 Spectral alignment and protocol-specific composition

The two protocols use the same spectral–sparse decomposition differently. In the Shared-Backbone Protocol, the spectral cores are consolidated into one shared update. Let $W(\cdot)$ denote the spectral-core whitening/alignment operation used to reduce pairwise coherence among retained low-rank

components, following prior spectral model-merging work (Gargiulo et al., 2025). Ada-HiSS applies W only to the low-rank cores and forms the shared spectral vector

$$\Delta^{(\ell)}_{\text{shared}} = \sum_{t=1}^{T} W(\Delta^{(\ell)}_{t,\text{core}}). \tag{18}$$

Sparse tails are deliberately left out of this shared sum: a tail is a task-specific patch activated only for the corresponding task, not a component merged into every evaluation model. At evaluation time, the task-specific vector is

$$\Delta_{\text{Ada-HiSS-Merge}}(t) = \alpha\, \Delta_{\text{shared}} + \lambda\, \Delta^{\text{SB}}_{t,\text{tail}}, \qquad \alpha = 0.9,\ \lambda = 1.0. \tag{19}$$

In the Task-Update Compression Protocol, no cross-task aggregation is performed: each task update is stored as a compact spectral–sparse package,

$$\widehat{\Delta}^{(\ell)}_t = \Delta^{(\ell)}_{t,\text{core}} + \Delta^{(\ell)}_{t,\text{tail}}, \tag{20}$$

and evaluated for task t by composing it with the shared base model,

$$\widehat{\theta}^{(\ell)}_t = \theta^{(\ell)}_0 + \widehat{\Delta}^{(\ell)}_t. \tag{21}$$

The decomposition clarifies why the method is useful in the Small-Room regime. The globally allocated core preserves the dominant low-rank task structure, and the alignment step reduces coherent low-rank interference before shared consolidation. The sparse tail preserves high-magnitude localized residual edits without forcing all task-specific corrections into the same shared vector. In shared-backbone composition, only the active task tail is patched onto the shared spectral vector; in task-update compression, the same decomposition yields a compact per-task replacement for the full task vector.

The spectral core is the main capacity allocator: it chooses where to spend low-rank degrees of freedom across the entire collection of task-layer matrices. The sparse tail is a bounded correction mechanism, preserving a small number of high-magnitude residual edits without changing the fact that the merge is primarily spectral.

4.5 Evaluation protocols

Shared-Backbone Protocol. This protocol targets deployment-oriented consolidation. For each task, we construct the task vector from a common initialization to the corresponding fine-tuned expert. Ada-HiSS pools singular values across matrix-shaped task-vector layers and tasks, selects a variable number of components using global thresholding, and applies spectral whitening/alignment to the retained cores. The aligned spectral components are aggregated into the single shared merged vector Δ_{shared} in Eq. (18).

The sparse part is handled as task-specific state. For task t, Ada-HiSS computes the residual between the task vector and the shared vector and stores only the top 1% residual coordinates by absolute magnitude, denoted $\Delta^{\text{SB}}_{t,\text{tail}}$. At evaluation time, the task-specific vector is composed as in Eq. (19). Thus, the Shared-Backbone Protocol is not a strict single-checkpoint merge: it stores one shared merged backbone plus a very small sparse residual tail per task. This is the intended trade-off. The shared spectral core handles global consolidation, while the sparse tail restores task-specific residual corrections without requiring task training data at merge time.

Task-Update Compression Protocol. This protocol isolates the quality of the Ada-HiSS task-update representation under a compact storage budget. Each compressed update $\widehat{\Delta}_t$ is stored separately for task t and applied to the base model only when evaluating that task, as in Eq. (21). This setting is useful for comparing against compact task-vector and coordinate-space baselines, because all methods operate directly on task updates under controlled storage budgets.

Detailed pseudocode for the two protocol instantiations is provided in Appendix B: Algorithm 1 gives the Shared-Backbone Protocol and Algorithm 2 gives the Task-Update Compression Protocol.

Table 1: **Evaluation protocols.** We separate shared-backbone consolidation from compact task-update storage to avoid conflating two deployment settings. Both protocols are data-free and use task-specific heads.

Protocol	Evaluation object	Backbone family	Baselines	Purpose
Shared-Backbone	A single merged spectral backbone plus a top-1% task-specific residual tail for Ada-HiSS.	OpenCLIP ViT-B/16, dense and Taylor-pruned	TSVM	Deployment-oriented consolidation under shared-backbone storage.
Task-Update Compression	A compact spectral–sparse update stored independently for each task; no cross-task aggregation.	Structured-pruned timm ViT-B/16	TIES, EMR, TSVC	Compact storage of per-task updates under a fixed budget.

5 EXPERIMENTS

5.1 SETUP

Backbone and pruning. The Shared-Backbone Protocol uses an OpenCLIP ViT-B/16 model family. We report the corresponding full dense model family and three recovered Taylor-pruned model families, with pruning ratios 0.3, 0.5, and 0.7. In all regimes, task experts are fine-tuned from the same shared initialization within that regime, and model merging uses only the resulting model parameters and task vectors. Appendix E reports the corresponding CLIP image/text parameter counts and sparse-tail storage overhead.

The Task-Update Compression Protocol uses a structured-pruned timm ViT-B/16-family backbone. The dense model has 86.57M parameters; the pruned backbone has 22.05M parameters, corresponding to approximately 74.5% parameter reduction. After pruning, the model is fine-tuned on ImageNet-1K to recover general-purpose features, yielding the shared pruned initialization θ_0. Each downstream expert is then fine-tuned from the same θ_0 with the pruned topology fixed.

Task suites. We evaluate 8-, 14-, and 20-task vision classification suites. The 8-task suite contains EuroSAT, DTD, Cars, SUN397, SVHN, RESISC45, MNIST, and GTSRB. The 14-task suite adds FER2013, PCAM, CIFAR100, Flowers102, Oxford-IIIT Pets, and STL10. The 20-task suite further adds KMNIST, EMNIST, RenderedSST2, FashionMNIST, Food101, and CIFAR10.

Baselines. We use TSVC to denote the task-update compression variant of task singular vectors, and TSVM to denote the shared-backbone merging variant used in the OpenCLIP protocol. In the Shared-Backbone Protocol, we compare against TSVM, a shared spectral merge baseline that produces one merged task vector/backbone using spectral orthogonalization. In the Task-Update Compression Protocol, we compare against TIES-Merging (Yadav et al., 2023), EMR-Merging (Huang et al., 2024), and TSVC (Gargiulo et al., 2025). All methods are evaluated with task-specific heads because the datasets have different label spaces.

Hyperparameters. In the Shared-Backbone Protocol, both the spectral baseline and Ada-HiSS use a fixed merge scale $\alpha = 0.9$, selected from the validation sweep in Appendix F.1. The baseline uses $k = 100$. Ada-HiSS uses global-threshold rank allocation with minimum rank one, a top-1% sparse residual tail per task, and tail scale $\lambda = 1.0$. In the Task-Update Compression Protocol, the main Ada-HiSS runs use a total compression budget of $k = 0.25$, with $p = 0.05$ reserved for the sparse residual tail and the remaining budget allocated to the spectral core; Appendix F.2 reports component and sparse-budget ablations.

Evaluation protocol and claim scope. Both protocols are data-free at construction time: they use the shared initialization and expert parameters/task vectors, but not training examples. We report top-1 accuracy averaged uniformly across tasks. Expert-normalized top-1 is computed as the merged-model top-1 accuracy divided by the corresponding single-task expert top-1 accuracy, averaged uniformly across tasks. In the Shared-Backbone Protocol, TSVM evaluates one shared merged backbone, while Ada-HiSS evaluates one shared merged spectral backbone plus the corresponding task's sparse

Table 2: **Shared-Backbone Protocol results on OpenCLIP ViT-B/16.** Average test accuracy and expert-normalized accuracy on full dense and Taylor-pruned OpenCLIP task suites. Ada-HiSS uses a shared spectral backbone with a top-1% sparse residual tail per task, improving retention in both capacity-constrained and dense regimes.

		Average top-1 (%)			Normalized top-1 (%)		
Regime	Method	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
Full dense	TSVM	82.90	80.61	78.81	88.87	87.22	84.74
	Ada-HiSS	**85.67**	**83.27**	**81.70**	**92.35**	**90.47**	**88.11**
Pruned 0.3	TSVM	80.56	77.92	77.08	87.06	84.96	83.40
	Ada-HiSS	**83.28**	**80.47**	**80.00**	**90.57**	**88.16**	**86.84**
Pruned 0.5	TSVM	76.16	73.19	72.36	82.92	80.42	78.99
	Ada-HiSS	**79.43**	**76.65**	**76.21**	**87.19**	**84.74**	**83.53**
Pruned 0.7	TSVM	66.40	64.89	64.02	72.80	72.28	71.43
	Ada-HiSS	**70.42**	**68.31**	**67.92**	**78.22**	**76.67**	**76.09**

Table 3: **Task-update compression results on structured-pruned timm ViT-B/16.** Average top-1 accuracy (%) is reported for compact task-update storage. Ada-HiSS improves over the spectral TSVC baseline and substantially outperforms coordinate-space baselines in the capacity-constrained regime.

Method	8 tasks	14 tasks	20 tasks
TIES (Yadav et al., 2023)	14.37	15.99	15.38
EMR (Huang et al., 2024)	62.97	59.02	51.50
TSVC (Gargiulo et al., 2025)	75.19	77.13	77.68
Ada-HiSS (ours)	**77.17**	**77.88**	**78.29**

residual tail. In the Task-Update Compression Protocol, each method stores a compact task update per task and evaluates θ_0 plus the corresponding update. The task heads remain task-specific for both protocols because the datasets have different label spaces.

5.2 Shared-Backbone Protocol results

The Shared-Backbone Protocol tests whether Ada-HiSS can improve a shared merged backbone while adding only compact task-specific residual tails.

Table 2 shows that Ada-HiSS consistently improves retention in dense and pruned regimes. In the full dense setting, Ada-HiSS improves average top-1 by 2.77, 2.66, and 2.89 points, with normalized gains of 3.48, 3.25, and 3.37 points. In the recovered pruned 0.3 setting, the corresponding top-1 gains are 2.72, 2.55, and 2.92 points, with normalized gains of 3.51, 3.20, and 3.44 points. In the recovered pruned 0.5 setting, the top-1 gains are 3.27, 3.46, and 3.85 points, with normalized gains of 4.27, 4.32, and 4.54 points. In the more aggressively pruned 0.7 setting, the top-1 gains are 4.02, 3.42, and 3.90 points, with normalized gains of 5.42, 4.39, and 4.66 points. The dense backbone retains the highest absolute accuracy, pruned 0.3 remains closest to the dense model, and pruned 0.7 shows the strongest Small-Room stress. Appendix E reports validation summaries, tail-storage overhead, and 20-task per-dataset deltas.

5.3 Task-Update Compression Protocol results

Table 3 shows that coordinate-wise parameter-space representations perform poorly in the pruned compression setting: TIES obtains only 14–16% average accuracy as the number of represented tasks increases. EMR is substantially stronger, but still below the spectral methods. TSVC is robust, indicating that spectral structure is valuable for compactly storing task updates in pruned models.

Table 4: **Interference-sensitive tasks under timm ViT-B/16 compression.** Accuracy (%) is reported for Cars and SUN397, fine-grained/scene-recognition tasks where interference is pronounced. Ada-HiSS gives larger gains over TSVC on these tasks than in the overall average.

	8 tasks		14 tasks		20 tasks	
Method	Cars	SUN397	Cars	SUN397	Cars	SUN397
TIES (Yadav et al., 2023)	2.85	8.07	1.42	5.04	0.89	2.19
EMR (Huang et al., 2024)	29.09	25.79	13.60	18.36	2.96	5.49
TSVC (Gargiulo et al., 2025)	58.97	34.37	63.84	36.79	63.85	36.78
Ada-HiSS (ours)	**67.50**	**39.63**	**67.63**	**39.67**	**67.81**	**39.84**

Table 5: **Additional pruned-DeiT compression check.** Average 20-task top-1 accuracy (%) is reported for tuned isomorphic-pruned DeiT models. Full 8-, 14-, and 20-task results, normalized accuracies, and model-size accounting are in Appendix G.

Model	TIES	EMR	TSVC	Ada-HiSS (ours)
deit_4.2G	32.11	73.87	85.78	**85.97**
deit_2.6G	23.46	69.10	85.95	**86.23**
deit_1.2G	20.70	56.09	80.81	**81.00**
deit_0.6G	16.41	56.71	79.71	**79.96**

Ada-HiSS improves over TSVC at every scale. The average gains are moderate, but consistent; the larger benefit appears on tasks where low-rank truncation alone loses localized task-specific information.

The strongest claim supported by Table 3 is therefore not that Ada-HiSS dramatically changes the mean over TSVC, but that a spectral representation is essential in the pruned regime and that the proposed hybrid allocation gives a small, repeatable improvement on top of that strong baseline.

Table 4 highlights the main empirical advantage of the hybrid representation. On Cars, Ada-HiSS improves over TSVC by 8.53, 3.79, and 3.96 points in the 8-, 14-, and 20-task suites, respectively. On SUN397, the corresponding improvements are 5.26, 2.88, and 3.06 points. This supports the interpretation that the sparse residual tail recovers fine-grained information that a purely low-rank merge can discard.

As an additional architecture check, we also evaluate Task-Update Compression on a tuned isomorphic-pruned DeiT family. Table 5 reports the 20-task setting across four pruning capacities. The result is intentionally compact: it is not used to tune Ada-HiSS, but it tests whether the spectral–sparse advantage persists beyond the main timm ViT-B/16 backbone. Ada-HiSS is consistently best across capacities, with the largest gaps against coordinate-space baselines and smaller but repeatable gains over TSVC.

Ablations for both protocols are reported in Appendix F. We keep them in the supplement because they support hyperparameter and component choices rather than changing the main empirical comparison.

6 Conclusion

We studied model merging in the Small-Room regime induced by structured pruning. Ada-HiSS addresses this regime by separating each task update into a globally allocated spectral component and a compact sparse residual component. Under the Shared-Backbone Protocol, this yields one shared merged spectral backbone plus a small task-specific residual tail, improving retention across dense and pruned multi-task suites. Under the Task-Update Compression Protocol, the same decomposition stores each task update compactly and explains why residual sparse corrections help most on interference-sensitive tasks. These results suggest that spectral allocation is a strong foundation for merging after pruning, and that small residual tails can recover task-specific information lost by a purely shared spectral update.

AI USE STATEMENT

We used AI writing assistants to polish the text and to port the paper to the ICLR LaTeX template. The method, experiments, and results are our own. We reviewed every AI-assisted edit and take full responsibility for the paper.

ETHICS STATEMENT

Our experiments use only public image-classification benchmarks (Appendix C) and involve no human subjects or personal data. We do not release new datasets or models.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix B describe the method in full, and Appendix A contains all proofs. Pruning, recovery, and fine-tuning recipes, dataset splits, and merge hyperparameters are listed in Appendix C; hyperparameter selection is covered in Appendix F.

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://openreview.net/forum?id=CQsmMYmlP5T.

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. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734.

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

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

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: An extension of mnist to handwritten letters, 2017.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16091–16101, 2023.

Gongfan Fang, Xinyin Ma, Michael Bi Mi, and Xinchao Wang. Isomorphic pruning for vision models. In *European Conference on Computer Vision*, pp. 232–250, 2024. doi: 10.1007/978-3-031-73404-5_14.

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

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

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017. doi: 10.1109/ICCV.2017.155.

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

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 122741–122769. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dda5cac5272a9bcd4bc73d90bc725ef1-Paper-Conference.pdf.

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://openreview.net/forum?id=6t0Kwf8-jrj.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

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. doi: 10.1109/ICCVW.2013.77.

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. doi: 10.1109/5.726791.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2736–2744, 2017. doi: 10.1109/ICCV.2017.298.

Michael S. Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716, 2022.

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 *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

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. doi: 10.1109/CVPR.2012.6248092.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. doi: 10.18653/v1/D13-1170. URL https://aclanthology.org/D13-1170/.

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.

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

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50268–50287. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24k.html.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms, 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 Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1644c9af28ab7916874f6fd6228a9bcf-Paper-Conference.pdf.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gMCsAqY7.

Mingyang Zhang, Jing Liu, Ganggui Ding, Linlin Ou, Xinyi Yu, and Bohan Zhuang. Channel merging: preserving specialization for merged experts. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i21.34405. URL https://doi.org/10.1609/aaai.v39i21.34405.

APPENDIX CONTENTS

- Appendix A: proofs of the Small-Room model and Ada-HiSS reconstruction statements.
- Appendix B: protocol-level pseudocode for shared-backbone merging and task-update compression.
- Appendix C: implementation details, ImageNet recovery, downstream fine-tuning, dataset splits, and dataset references.
- Appendix D: additional discussion and limitations.
- Appendix E: OpenCLIP shared-backbone storage accounting and supplementary results.
- Appendix F: ablations for the Shared-Backbone and Task-Update Compression protocols.
- Appendix G: full supplementary DeiT comparison.
- Appendix H: dense-versus-pruned TIES diagnostic.
- Appendix I: per-dataset task-update compression results.

A PROOFS AND ADDITIONAL DETAILS

This appendix gives compact proofs for the proposition statements in the main paper, followed by an expansion of the local interference objective. The results are reconstruction or diagnostic statements: they support the design of Ada-HiSS, but they do not assert a downstream accuracy optimality guarantee.

Proof of Proposition 1. Let $P_{\mathcal{U}}$ denote the orthogonal projector onto the r-dimensional subspace $\mathcal{U}$. Condition on z_t. For each $j \neq t$, isotropy of the unit vector z_j in $\mathcal{U}$ gives

$$\mathbb{E}[z_j] = 0, \qquad \mathbb{E}[z_j z_j^\top] = \frac{1}{r} P_{\mathcal{U}}. \tag{22}$$

Define $v = \sum_{j \neq t} z_j$. By independence and zero mean, the cross terms vanish, so

$$\mathbb{E}[vv^\top] = \sum_{j \neq t} \mathbb{E}[z_j z_j^\top] = \frac{T-1}{r} P_{\mathcal{U}}. \tag{23}$$

Therefore,

$$\mathbb{E}[I_t \mid z_t] = \mathbb{E}\left[z_t^\top v v^\top z_t \mid z_t\right] \tag{24}$$

$$= z_t^\top \left(\frac{T-1}{r} P_{\mathcal{U}}\right) z_t = \frac{T-1}{r}, \tag{25}$$

where the last equality uses $z_t \in \mathcal{U}$ and $\|z_t\|_2 = 1$. Taking expectation over z_t gives $\mathbb{E}[I_t] = (T-1)/r$. Under the same normalized isotropic model and the same number of tasks, replacing r_{dense} by r_{pruned} changes the expectation by the factor

$$\frac{(T-1)/r_{\text{pruned}}}{(T-1)/r_{\text{dense}}} = \frac{r_{\text{dense}}}{r_{\text{pruned}}}. \tag{26}$$

Proof of Proposition 2. For a fixed matrix A_m, the rank-one SVD components $u_{m,i} v_{m,i}^\top$ are orthonormal under the Frobenius inner product. Thus, for any selection variables $x_{m,i} \in \{0, 1\}$,

$$\|A_m - \widehat{A}_m\|_F^2 = \sum_i (1 - x_{m,i}) \sigma_{m,i}^2. \tag{27}$$

Summing over all task-layer matrices gives

$$\sum_m \|A_m - \widehat{A}_m\|_F^2 = \sum_{m,i} \sigma_{m,i}^2 - \sum_{m,i} x_{m,i} \sigma_{m,i}^2. \tag{28}$$

The first term is constant, so minimizing the reconstruction error is equivalent to maximizing

$$\sum_{m,i} x_{m,i} \sigma_{m,i}^2 \qquad \text{subject to} \qquad \sum_{m,i} x_{m,i} \leq K. \tag{29}$$

One maximizer sets $x_{m,i} = 1$ for the K largest values of $\sigma_{m,i}^2$, equivalently the K largest singular values, with arbitrary tie-breaking; if fewer than K singular values are nonzero, adding zero modes does not change the objective. This proves the claim. If a minimum-rank-one implementation guard is imposed, the mandatory top component of each eligible matrix is reserved first, and the same argument applies to the remaining optional components under the remaining budget.

Proof of Proposition 3. Fix a support set Ω with $|\Omega| \le s$. Since the Frobenius objective decomposes entrywise, the best matrix Z supported on Ω satisfies

$$Z_{ij} = R_{ij} \quad \text{for } (i,j) \in \Omega, \qquad Z_{ij} = 0 \quad \text{for } (i,j) \notin \Omega. \tag{30}$$

For this support, the remaining error is

$$\|R - Z\|_F^2 = \sum_{(i,j)\notin\Omega} R_{ij}^2. \tag{31}$$

Therefore, minimizing the error is equivalent to choosing a support Ω that maximizes the retained energy

$$\sum_{(i,j)\in\Omega} R_{ij}^2. \tag{32}$$

The optimal support is the set of s entries with largest squared magnitude, equivalently largest absolute value. Thus one optimal solution is $Z = \mathcal{T}_s(R)$, with arbitrary tie-breaking. If the sparse budget is applied globally across several residual matrices, the same entrywise proof applies after vectorizing the collection of residuals.

From quadratic error to interference. Expanding Eq. (4) gives

$$\|(\alpha-1)\Delta_t\|_{H_t}^2 + \alpha^2 \left\| \sum_{j\neq t} \Delta_j \right\|_{H_t}^2 + 2\alpha(\alpha-1) \left\langle \Delta_t, \sum_{j\neq t} \Delta_j \right\rangle_{H_t}.$$

The first term is shrinkage/overshoot of the target task update, while the second and third terms are cross-task interference. In a pruned model, H_t is evaluated over a reduced architecture whose remaining directions are more likely to be shared by multiple tasks. This is the formal sense in which pruning turns harmless redundancy into collision.

B PROTOCOL ALGORITHMS

Algorithm 1: Ada-HiSS Shared-Backbone Protocol

Input: shared initialization θ_0, task experts $\{\theta_t\}_{t=1}^T$, matrix-layer set $\mathcal{M}$, spectral budget K or fraction k, merge scale α, sparse-tail budget p, tail scale λ.
Output: one shared spectral vector Δ_{shared} and task-specific sparse tails $\{\Delta_{t,\text{tail}}^{\text{SB}}\}_{t=1}^T$.

1. Compute task updates $\Delta_t = \theta_t - \theta_0$ for all tasks t.
2. For each t and each matrix layer $\ell \in \mathcal{M}$, compute the compact SVD of $\Delta_t^{(\ell)}$ as in Eq. (7); collect all singular values in $\mathcal{S}$.
3. Choose the global threshold τ that retains the top K singular values in $\mathcal{S}$, optionally enforcing the minimum-rank guard used in implementation.
4. Reconstruct the spectral core $\Delta_{t,\text{core}}^{(\ell)}$ for each task-layer pair using Eq. (9).
5. Apply the spectral alignment operator $W(\cdot)$ to the retained cores and form the shared spectral vector $\Delta_{\text{shared}}^{(\ell)} = \sum_{t=1}^T W(\Delta_{t,\text{core}}^{(\ell)})$ for every matrix layer ℓ.
6. For each task t, compute the shared-backbone residual $R_t^{\text{SB}} = \Delta_t - \Delta_{\text{shared}}$ and store the sparse task tail $\Delta_{t,\text{tail}}^{\text{SB}} = \mathcal{T}_p(R_t^{\text{SB}})$.
7. At evaluation time for task t, compose the active task vector $\Delta_{\text{Ada-HiSS-Merge}}(t) = \alpha\Delta_{\text{shared}} + \lambda\Delta_{t,\text{tail}}^{\text{SB}}$ and evaluate $\theta_0 + \Delta_{\text{Ada-HiSS-Merge}}(t)$ with task head h_t.

Note: the construction is data-free and stores one shared backbone plus compact task-specific residual state.

Algorithm 2: Ada-HiSS Task-Update Compression Protocol

Input: shared initialization θ_0, task experts $\{\theta_t\}_{t=1}^{T}$, matrix-layer set $\mathcal{M}$, spectral budget K or fraction k, sparse-tail budget p or per-layer counts $\{s_\ell\}$.

Output: compact per-task spectral–sparse updates $\{\widehat{\Delta}_t\}_{t=1}^{T}$.

1. Compute task updates $\Delta_t = \theta_t - \theta_0$ for all tasks t.
2. For each t and matrix layer $\ell \in \mathcal{M}$, compute the compact SVD of $\Delta_t^{(\ell)}$ and collect the singular values across all tasks and matrix layers.
3. Choose the global threshold τ that retains the top K singular values, and reconstruct each spectral core $\Delta_{t,\mathrm{core}}^{(\ell)}$ using Eq. (9).
4. For each task-layer pair, compute the compression residual $R_t^{(\ell)} = \Delta_t^{(\ell)} - \Delta_{t,\mathrm{core}}^{(\ell)}$ and retain the sparse tail $\Delta_{t,\mathrm{tail}}^{(\ell)} = \mathcal{T}_{s_\ell}(R_t^{(\ell)})$.
5. Store each task package as retained spectral factors for $\Delta_{t,\mathrm{core}}^{(\ell)}$ plus sparse coordinates and values for $\Delta_{t,\mathrm{tail}}^{(\ell)}$ across layers.
6. At evaluation time for task t, reconstruct $\widehat{\Delta}_t^{(\ell)} = \Delta_{t,\mathrm{core}}^{(\ell)} + \Delta_{t,\mathrm{tail}}^{(\ell)}$ and evaluate $\widehat{\theta}_t = \theta_0 + \widehat{\Delta}_t$ with task head h_t.

Note: no cross-task aggregation is performed; each task update is stored and applied independently under the compact budget.

C Additional experimental details

Pruning pipeline. For the OpenCLIP shared-backbone experiments, the dense CLIP ViT-B/16 image encoder is structurally pruned and then recovered on ImageNet-1K before downstream task fine-tuning. The topology is fixed for all downstream experts. No additional pruning is applied during downstream fine-tuning or merging.

For the OpenCLIP merged-backbone runs, we use the dense OpenCLIP ViT-B/16 model and recovered Taylor-pruned variants at pruning ratios 0.3, 0.5, and 0.7. Downstream task experts are fine-tuned separately for each dataset from the corresponding shared initialization.

OpenCLIP ImageNet recovery. After structural pruning, each pruned OpenCLIP image encoder is recovered by distilling features from the frozen dense CLIP ViT-B/16 image encoder. The teacher is the original dense `ViT-B-16 ImageEncoder` with `keep_lang=False`; the student is the corresponding structurally pruned image encoder. Recovery uses ImageNet-1K train images (1,281,167 examples) with `torchvision.datasets.ImageFolder` and the recovered student's CLIP train/validation preprocessors. We train for 5 epochs with AdamW, learning rate 10^{-5}, weight decay 0.05, batch size 1024, validation batch size 512, 500 warmup steps followed by cosine decay, gradient clipping at 1.0, and `bf16` mixed precision. The recovery objective is

$$\mathcal{L}_{\mathrm{rec}} = 1 - \cos(\bar{f}_s(x), \bar{f}_t(x)) + 0.25\|\bar{f}_s(x) - \bar{f}_t(x)\|_2^2, \tag{33}$$

where $\bar{f}_s(x)$ and $\bar{f}_t(x)$ are L2-normalized student and teacher image features. The teacher is frozen and evaluated under `torch.no_grad()`.

Table 6: **OpenCLIP ImageNet feature recovery.** Validation is measured on 20 ImageNet validation batches, i.e. 10,240 images.

Recovered model	Final recovery loss	Final feature cosine
Taylor 0.3	0.020478	0.979542
Taylor 0.5	0.037194	0.962842
Taylor 0.7	0.062019	0.938042

OpenCLIP downstream task-expert fine-tuning. After recovery, each OpenCLIP model is fine-tuned independently on every downstream task. The same fine-tuning procedure and hyperparameters are used for full dense CLIP and all recovered pruned models, differing only in the starting model. We use the task-specific CLIP classification head for the corresponding `DatasetVal` task and freeze the head, updating only the image encoder. Each run uses AdamW with learning rate 10^{-5}, weight decay 0.1, batch size 128, 500 warmup steps followed by cosine decay, cross-entropy loss, label smoothing 0.0, gradient clipping at 1.0, and `bf16` mixed precision. Fine-tuning jobs were run in parallel on a single RTX PRO 6000 Blackwell Max-Q GPU, with up to 8 concurrent pruned-model jobs and up to 4 concurrent full dense CLIP jobs to avoid memory pressure.

Task-expert fine-tuning uses the `DatasetVal` wrapper. After fine-tuning, we evaluate each expert on both `DatasetVal` and the corresponding held-out `Dataset` test wrapper; main paper tables use held-out test top-1 unless explicitly marked as validation.

Table 7: **OpenCLIP downstream fine-tuning schedule.** Epoch counts are task-specific; all other optimization settings are shared across dense and recovered pruned OpenCLIP models.

Dataset	Epochs	Dataset	Epochs
Cars	35	STL10	60
DTD	76	OxfordIIITPet	82
EuroSAT	12	Flowers102	147
GTSRB	11	CIFAR100	6
MNIST	5	PCAM	1
RESISC45	15	FER2013	10
SUN397	14	CIFAR10	6
SVHN	4	Food101	4
FashionMNIST	5	RenderedSST2	39
EMNIST	2	KMNIST	5

Shared-Backbone Protocol implementation details. The shared spectral baseline uses TSVM with rank budget $k = 100$, without TIES-style sign aggregation. Ada-HiSS uses global-threshold rank allocation with the automatic global-rank fraction, minimum rank one, a top-1% sparse residual tail per task, and tail scale $\lambda = 1.0$. The fixed-alpha comparison uses $\alpha = 0.9$ for all rows. At evaluation time, the active task tail is added to the shared spectral vector before applying the task-specific head.

All Shared-Backbone Protocol merge operations are data-free at merge time: they use model parameters and task vectors, not task training images. All reported results use task-specific heads because the datasets have different label spaces. Average top-1 values are macro-averages over tasks. Table 8 reports deployment-relevant parameter counts including the unchanged CLIP text tower.

Validation and test splits. The reported test numbers use each dataset wrapper's held-out `test_dataset`. For standard datasets such as CIFAR10, CIFAR100, MNIST, SVHN, GTSRB, Cars, Oxford-IIIT Pets, Flowers102, Food101, PCAM, STL10, KMNIST, and EMNIST, this corresponds to the official test split provided by the benchmark. For datasets distributed through custom folders or split files, such as EuroSAT, RESISC45, DTD, and SUN397, the wrapper treats the designated held-out folder or split file as the test split. Validation numbers use the corresponding `DatasetVal` wrapper: if an explicit validation split exists, the wrapper uses it; otherwise it constructs validation data by holding out 10% of the training set, capped at 5000 validation examples. Thus final test accuracy is evaluated on a split separate from validation selection. Some raw benchmark layouts name their held-out evaluation folder "val", but those splits are used only for held-out evaluation.

We use TSVC to denote the Task-Update Compression Protocol spectral baseline and TSVM to denote the Shared-Backbone Protocol spectral baseline.

Dataset references. The task suite includes EuroSAT (Helber et al., 2019), DTD (Cimpoi et al., 2014), Cars (Krause et al., 2013), SUN397 (Xiao et al., 2010), SVHN (Netzer et al., 2011), RESISC45 (Cheng et al., 2017), MNIST (LeCun et al., 1998), GTSRB (Stallkamp et al., 2011), FER2013 (Goodfellow et al., 2015), PCAM (Veeling et al., 2018), CIFAR10/100 (Krizhevsky, 2009), Flowers102

(Nilsback & Zisserman, 2008), Oxford-IIIT Pets (Parkhi et al., 2012), STL10 (Coates et al., 2011), KMNIST (Clanuwat et al., 2018), EMNIST (Cohen et al., 2017), Fashion-MNIST (Xiao et al., 2017), Food101 (Bossard et al., 2014), and RenderedSST2 derived from SST (Socher et al., 2013).

D Additional discussion and limitations

The evidence supports a focused claim: Ada-HiSS improves data-free consolidation and compact storage of task updates in multi-dataset vision classification when capacity is constrained by structured pruning. The Shared-Backbone Protocol evaluates deployment-oriented consolidation on dense and recovered Taylor-pruned OpenCLIP ViT-B/16 model families, while the Task-Update Compression Protocol evaluates compact per-task update storage on a structured-pruned timm ViT-B/16 backbone. These protocols answer related but distinct questions, so their results should be read as complementary rather than interchangeable.

The Shared-Backbone Protocol changes the deployment profile relative to strict single-checkpoint merging: Ada-HiSS stores one shared backbone plus a sparse residual tail per task. We therefore report tail-storage overhead in Appendix E and present the method as a storage–accuracy trade-off. These tails are small compared with storing one full expert per task, but they are task-specific state.

Several limitations remain. All reported results use task-specific heads, which isolates backbone merging but does not test whether a single shared classifier can be merged. The current experiments use ViT-B/16-family backbones and single reported runs rather than a broad multi-seed study. Additional architectures, pruning recipes, latency measurements, and broader task families such as language, detection, or segmentation remain important future tests. Finally, the merge operation is data-free at merge time because it operates on model parameters and task updates; hyperparameter selection should be reported separately whenever validation accuracy is used.

E Shared-Backbone Protocol details

This section records additional Shared-Backbone Protocol outputs. The main text reports test averages for full dense, pruned 0.3, pruned 0.5, and pruned 0.7 model families; the supplementary tables below provide model-size accounting, Ada-HiSS tail-storage overhead, validation averages, and per-dataset 20-task deltas.

Table 9 quantifies the extra task-specific state used by Ada-HiSS-Merge. With the current top-1% tail density and a flat `int32` index plus `fp32` value representation, each task tail costs about 2% of the shared backbone size. The total tail overhead is therefore about 16%, 28%, and 40% of the shared backbone for 8-, 14-, and 20-task suites, respectively; using `fp16` tail values reduces these overheads to about 12%, 21%, and 30%. These tails are much smaller than storing one complete expert per task, but they are reported explicitly because Ada-HiSS-Merge is a shared-backbone-plus-tail method rather than a strict single-checkpoint merge.

F Ablation studies

F.1 Shared-Backbone Protocol ablations

This subsection reports ablations for Ada-HiSS-Merge under the Shared-Backbone Protocol. Unless stated otherwise, these runs use global-threshold rank allocation with minimum rank one, a top-1% sparse residual tail per task, tail scale $\lambda = 1.0$, and task-specific frozen heads. These ablations are used only to justify fixed hyperparameter choices; the main claims use the held-out test results in Table 2.

The automatic global-rank rule keeps the absolute shared spectral budget approximately constant as the number of merged tasks changes. For the 8-, 14-, and 20-task suites, the target shared budget is 38,162 singular directions, corresponding to 12.5%, 7.14%, and 5.0% of the multi-task spectral pool, respectively. Table 15 reports the resulting allocation diagnostics for the most aggressively pruned OpenCLIP setting.

Table 8: **Model size after structured pruning.** Current rows report OpenCLIP ViT-B/16 parameter counts including the pruned CLIP image tower and the unchanged CLIP text tower. Total reduction is measured against the full dense CLIP model. FP32 storage is the parameter-storage equivalent for the full deployed CLIP model; serialized storage is smaller because the full text transformer is omitted.

Model	Image params	Text params	Total params	Total reduction	FP32 storage	Serialized size
Full dense CLIP ViT-B/16	86.19M	63.43M	149.62M	0.00%	570.76 MiB	426.56 MiB
Pruned Taylor 0.3	69.19M	63.43M	132.62M	11.37%	505.89 MiB	361.69 MiB
Pruned Taylor 0.5	57.86M	63.43M	121.29M	18.93%	462.69 MiB	318.49 MiB
Pruned Taylor 0.7	46.52M	63.43M	109.95M	26.52%	419.42 MiB	275.22 MiB

Table 9: **Ada-HiSS-Merge storage overhead.** The shared backbone is stored once, while each task adds a top-1% sparse residual tail. Tail storage assumes one flat `int32` index and one `fp32` value per nonzero; the `fp16` variant uses `fp16` values with `int32` indices.

Model	Tasks	Shared fp32	Tails fp32+i32	Tails fp16+i32
Full dense CLIP ViT-B/16	8	426.45 MiB	68.23 MiB	51.17 MiB
	14	426.45 MiB	119.41 MiB	89.56 MiB
	20	426.45 MiB	170.58 MiB	127.94 MiB
Pruned Taylor 0.3	8	361.58 MiB	57.85 MiB	43.39 MiB
	14	361.58 MiB	101.24 MiB	75.93 MiB
	20	361.58 MiB	144.63 MiB	108.47 MiB
Pruned Taylor 0.5	8	318.38 MiB	50.94 MiB	38.21 MiB
	14	318.38 MiB	89.15 MiB	66.86 MiB
	20	318.38 MiB	127.35 MiB	95.51 MiB
Pruned Taylor 0.7	8	275.11 MiB	44.02 MiB	33.01 MiB
	14	275.11 MiB	77.03 MiB	57.77 MiB
	20	275.11 MiB	110.04 MiB	82.53 MiB

Table 10: **Validation and test averages for OpenCLIP merging.** All rows use fixed merge scale $\alpha = 0.9$ and task-specific heads. Normalized accuracy uses each dataset's single-task expert as the reference. Bold entries mark the better method for each regime, task count, and metric.

Regime	Method	Tasks	Val. top-1	Test top-1	Val. norm.	Test norm.
Pruned 0.3	TSVM	8	80.50	80.56	86.92	87.06
	Ada-HiSS-Merge		**83.45**	**83.28**	**90.69**	**90.57**
	TSVM	14	78.45	77.92	84.65	84.96
	Ada-HiSS-Merge		**81.01**	**80.47**	**87.89**	**88.16**
	TSVM	20	77.28	77.08	83.35	83.40
	Ada-HiSS-Merge		**80.27**	**80.00**	**86.85**	**86.84**
Pruned 0.5	TSVM	8	75.83	76.16	82.70	82.92
	Ada-HiSS-Merge		**79.32**	**79.43**	**87.24**	**87.19**
	TSVM	14	73.77	73.19	80.04	80.42
	Ada-HiSS-Merge		**77.12**	**76.65**	**84.27**	**84.74**
	TSVM	20	72.93	72.36	79.31	78.99
	Ada-HiSS-Merge		**76.86**	**76.21**	**83.93**	**83.53**
Pruned 0.7	TSVM	8	65.89	66.40	72.17	72.80
	Ada-HiSS-Merge		**69.87**	**70.42**	**77.48**	**78.22**
	TSVM	14	65.08	64.89	71.64	72.28
	Ada-HiSS-Merge		**68.44**	**68.31**	**75.89**	**76.67**
	TSVM	20	64.56	64.02	71.99	71.43
	Ada-HiSS-Merge		**68.74**	**67.92**	**76.94**	**76.09**
Full dense	TSVM	8	82.44	82.90	88.52	88.87
	Ada-HiSS-Merge		**84.76**	**85.67**	**91.43**	**92.35**
	TSVM	14	80.65	80.61	86.57	87.22
	Ada-HiSS-Merge		**83.25**	**83.27**	**89.74**	**90.47**
	TSVM	20	78.97	78.81	84.73	84.74
	Ada-HiSS-Merge		**81.94**	**81.70**	**88.19**	**88.11**

Table 11: **Per-dataset 20-task deltas after aggressive pruning.** Δ is Ada-HiSS-Merge minus TSVM in percentage points for the recovered pruned 0.7 OpenCLIP setting. Bold entries mark the higher value in each TSVM/Ada pair.

Dataset	TSVM	Ada-HiSS-Merge	Δ	TSVM norm.	Ada norm.	Δ norm.
Cars	17.17	**25.77**	8.59	24.52	**36.78**	12.27
DTD	40.21	**45.21**	5.00	54.00	**60.71**	6.71
EuroSAT	78.15	**83.96**	5.81	78.91	**84.78**	5.87
GTSRB	66.32	**72.63**	6.31	66.94	**73.31**	6.37
MNIST	99.21	**99.27**	0.06	99.53	**99.59**	0.06
RESISC45	55.54	**64.33**	8.79	58.13	**67.34**	9.20
SVHN	83.57	**87.85**	4.27	85.63	**90.01**	4.38
SUN397	13.42	**20.36**	6.95	23.21	**35.23**	12.02
STL10	92.73	**93.05**	0.32	95.93	**96.26**	0.34
Oxford-IIIT Pet	84.25	**85.94**	1.69	92.21	**94.06**	1.85
Flowers102	41.88	**42.51**	0.63	48.08	**48.81**	0.73
CIFAR100	54.34	**61.75**	7.41	63.39	**72.04**	8.64
PCAM	**78.12**	77.32	-0.81	**89.63**	88.71	-0.92
FER2013	47.26	**50.31**	3.05	66.19	**70.46**	4.27
CIFAR10	90.15	**92.04**	1.89	92.47	**94.41**	1.94
Food101	50.90	**56.33**	5.44	59.58	**65.94**	6.37
FashionMNIST	81.99	**84.30**	2.31	86.21	**88.63**	2.43
RenderedSST2	55.57	**55.74**	0.16	93.53	**93.81**	0.28
EMNIST	99.44	**99.48**	0.05	99.65	**99.70**	0.05
KMNIST	50.21	**60.25**	10.04	50.95	**61.14**	10.19

Table 12: **Per-dataset 20-task deltas after mild pruning.** Δ is Ada-HiSS-Merge minus TSVM in percentage points for the recovered pruned 0.3 OpenCLIP setting. Bold entries mark the higher value in each TSVM/Ada pair.

Dataset	TSVM	Ada-HiSS-Merge	Δ	TSVM norm.	Ada norm.	Δ norm.
Cars	51.50	**60.76**	9.27	61.09	**72.09**	10.99
DTD	59.52	**64.10**	4.57	74.55	**80.28**	5.73
EuroSAT	93.96	**94.96**	1.00	94.49	**95.49**	1.01
GTSRB	81.43	**83.25**	1.83	82.25	**84.09**	1.85
MNIST	**99.42**	**99.42**	0.00	**99.71**	**99.71**	0.00
RESISC45	77.89	**82.89**	5.00	80.73	**85.92**	5.18
SVHN	87.48	**89.59**	2.11	89.47	**91.63**	2.16
SUN397	22.76	**35.07**	12.31	32.20	**49.62**	17.42
STL10	97.36	**97.64**	0.28	99.06	**99.34**	0.28
Oxford-IIIT Pet	92.04	**92.20**	0.16	98.11	**98.29**	0.17
Flowers102	**72.06**	70.06	-2.00	**77.37**	75.22	-2.15
CIFAR100	71.04	**75.82**	4.78	79.08	**84.40**	5.32
PCAM	82.60	**83.03**	0.44	92.92	**93.41**	0.49
FER2013	58.08	**60.92**	2.84	78.59	**82.43**	3.85
CIFAR10	94.80	**95.71**	0.91	96.40	**97.33**	0.93
Food101	80.77	**82.94**	2.17	88.37	**90.74**	2.37
FashionMNIST	85.40	**87.21**	1.81	89.56	**91.45**	1.90
RenderedSST2	73.92	**74.96**	1.04	93.34	**94.66**	1.32
EMNIST	99.57	**99.60**	0.04	99.75	**99.78**	0.04
KMNIST	59.94	**69.76**	9.82	60.88	**70.86**	9.97

Table 13: **Per-dataset 20-task deltas after moderate pruning.** Δ is Ada-HiSS-Merge minus TSVM in percentage points for the recovered pruned 0.5 OpenCLIP setting. Bold entries mark the higher value in each TSVM/Ada pair.

Dataset	TSVM	Ada-HiSS-Merge	Δ	TSVM norm.	Ada norm.	Δ norm.
Cars	34.03	**44.07**	10.05	42.79	**55.43**	12.64
DTD	51.12	**57.02**	5.90	65.20	**72.73**	7.53
EuroSAT	89.85	**91.67**	1.81	90.45	**92.28**	1.83
GTSRB	78.02	**80.66**	2.64	78.86	**81.53**	2.67
MNIST	99.35	**99.38**	0.03	99.63	**99.66**	0.03
RESISC45	68.79	**76.84**	8.05	71.62	**80.00**	8.38
SVHN	85.39	**88.47**	3.07	87.41	**90.56**	3.15
SUN397	19.57	**29.94**	10.37	29.44	**45.04**	15.60
STL10	95.83	**96.38**	0.55	97.84	**98.40**	0.56
Oxford-IIIT Pet	89.34	**90.24**	0.90	95.96	**96.93**	0.97
Flowers102	58.37	**59.64**	1.27	64.18	**65.58**	1.39
CIFAR100	66.28	**72.20**	5.92	75.22	**81.93**	6.72
PCAM	78.82	**80.64**	1.82	88.16	**90.20**	2.04
FER2013	54.69	**57.86**	3.16	75.33	**79.68**	4.36
CIFAR10	93.35	**94.56**	1.21	95.09	**96.32**	1.23
Food101	72.92	**76.80**	3.88	81.31	**85.64**	4.33
FashionMNIST	84.18	**86.20**	2.02	88.38	**90.50**	2.12
RenderedSST2	68.04	**70.02**	1.98	92.67	**95.36**	2.69
EMNIST	99.52	**99.56**	0.03	99.74	**99.77**	0.04
KMNIST	59.67	**72.13**	12.46	60.52	**73.16**	12.64

Table 14: **Per-dataset 20-task deltas in the dense setting.** Δ is Ada-HiSS-Merge minus TSVM in percentage points. Bold entries mark the higher value in each TSVM/Ada pair.

Dataset	TSVM	Ada-HiSS-Merge	Δ	TSVM norm.	Ada norm.	Δ norm.
Cars	65.20	**70.31**	5.11	74.78	**80.64**	5.86
DTD	62.23	**66.65**	4.41	75.48	**80.84**	5.35
EuroSAT	94.41	**95.56**	1.15	95.25	**96.41**	1.16
GTSRB	**83.12**	82.32	-0.80	**84.09**	83.28	-0.81
MNIST	**99.50**	**99.50**	0.00	**99.70**	**99.70**	0.00
RESISC45	83.83	**87.19**	3.37	86.70	**90.18**	3.48
SVHN	86.85	**89.11**	2.26	88.80	**91.12**	2.31
SUN397	25.95	**39.94**	13.98	34.56	**53.18**	18.62
STL10	97.82	**98.04**	0.21	98.91	**99.13**	0.21
Oxford-IIIT Pet	**93.19**	**93.19**	0.00	**98.87**	**98.87**	0.00
Flowers102	**79.96**	78.99	-0.98	**83.86**	82.84	-1.02
CIFAR100	75.07	**79.17**	4.10	82.94	**87.47**	4.53
PCAM	83.17	**85.27**	2.10	93.56	**95.92**	2.36
FER2013	58.39	**62.43**	4.04	78.29	**83.71**	5.42
CIFAR10	95.55	**96.20**	0.65	96.94	**97.60**	0.66
Food101	85.56	**87.17**	1.61	91.97	**93.70**	1.73
FashionMNIST	87.64	**88.93**	1.29	91.55	**92.90**	1.35
RenderedSST2	73.31	**75.67**	2.36	92.13	**95.10**	2.97
EMNIST	99.58	**99.63**	0.05	99.80	**99.85**	0.05
KMNIST	45.83	**58.77**	12.94	46.55	**59.70**	13.14

Table 15: **Global-rank allocation diagnostics for Ada-HiSS-Merge.** The table reports the global threshold and layer allocation statistics for the pruned Taylor 0.7 OpenCLIP setting. The target budget is held fixed while the number of task updates grows, so the effective fraction of retained spectral components decreases with the task count.

Tasks	Global threshold	Target budget	Allocation min	Allocation max	Allocation mean	Tail density
8	0.010704	38,162 / 305,296	8	1,495	734.19	1.0%
14	0.015038	38,162 / 534,268	14	1,615	734.42	1.0%
20	0.018469	38,162 / 763,240	20	1,687	734.65	1.0%

We select the Shared-Backbone merge coefficient on validation data using the 20-task pruned Taylor 0.7 setting, which is the most capacity-constrained OpenCLIP configuration. Table 16 shows a clear peak around $\alpha = 0.9$–1.0; performance then decreases steadily as the merge coefficient becomes larger. We therefore use $\alpha = 0.9$ for all fixed-alpha Shared-Backbone results rather than tuning it separately for each pruning ratio or task count.

Table 17 verifies that the selected coefficient transfers across dense and pruned OpenCLIP regimes and across 8-, 14-, and 20-task suites. Accuracy decreases as pruning becomes more aggressive, as expected, but the same Ada-HiSS-Merge hyperparameters remain usable without per-setting retuning.

These Shared-Backbone ablations isolate merge-scale selection and document the global-rank allocation used by Ada-HiSS-Merge. They do not isolate tail density, tail scale, or alternative rank-allocation rules under otherwise identical Shared-Backbone settings; those remain separate controlled ablations.

F.2 Task-Update Compression Protocol ablations

We ablate the main Task-Update Compression Protocol design choices on the 8-task suite, since this suite contains the two most interference-sensitive datasets (Cars and SUN397) while still covering diverse satellite, texture, traffic-sign, digit, and scene-recognition tasks. Unless stated otherwise, these runs use full SUN397 evaluation, $k = 0.25$, task-specific heads, and the same structured-pruned timm ViT-B/16 experts as in Table 3. We denote the sparse residual budget by p to avoid overloading the Shared-Backbone merge-scale notation α.

Table 18 shows three clear trends. First, the spectral core is the dominant component. Removing it and retaining only a sparse raw-update tail reduces average accuracy to 17.75%, confirming that the large task signal is not well captured by isolated coordinates. This supports the central premise of Ada-HiSS: in the Small-Room regime, the useful task update is structured and matrix-valued, not merely sparse in the original parameter basis.

Second, global spectral allocation matters. Replacing the global singular-value threshold with a fixed per-layer rank lowers average accuracy from 77.17% to 76.66%. The drop is especially visible on Cars and SUN397, where fixed rank loses 1.79 and 1.62 points, respectively, relative to the default. This matches Proposition 2: a global threshold spends rank where the spectra indicate useful energy, whereas a uniform per-layer rank can waste capacity on layers with weaker singular modes and under-allocate layers with sharper task-specific structure.

Third, the residual-tail choice is preferable to selecting sparse entries directly from the raw update. The raw-update tail remains competitive, but it is slightly worse than the residual tail in both average accuracy (77.12 vs. 77.17) and normalized accuracy (88.47 vs. 88.53). This supports the intended division of labor: the spectral core first captures smooth high-energy directions, and the sparse tail should then recover what the core missed rather than duplicate large entries already represented by the low-rank approximation.

The spectral core is necessary and nearly sufficient on this 8-task ablation: the default sparse tail changes the mean only slightly, and the core-only row is marginally higher on Cars and SUN397. We therefore interpret global spectral allocation as the main stabilizer, while the sparse residual tail is a bounded correction that is inexpensive, does not make performance brittle, and can recover small task-specific details in some settings.

Table 19 studies the sparse residual budget with k fixed. The curve is relatively flat between $p = 0.02$ and $p = 0.10$, suggesting that Ada-HiSS is not brittle to the exact tail budget in the low-sparsity region. Increasing p from 0.05 to 0.08 gives a small gain in the capped sweep and also gives the best full-SUN average in Table 18 (77.20%). The improvement is small, so we keep $p = 0.05$ as the default setting for the main comparison because it is simpler and matches the original fixed-budget decomposition, while reporting $p = 0.08$ as a sensitivity result.

This sensitivity diagnostic suggests that the sparse budget is not hand-tuned to a single lucky value; the main comparison uses one fixed default, while the sweep checks robustness around that choice.

Table 16: **Shared-Backbone merge-scale ablation.** Validation performance for Ada-HiSS-Merge on the 20-task pruned Taylor 0.7 OpenCLIP setting. We select $\alpha = 0.9$ and reuse it for all fixed-alpha test runs.

α	Avg. top-1	Avg. normalized top-1
0.5	63.99	71.85
0.7	67.38	75.52
0.8	68.33	76.53
0.9	**68.74**	**76.94**
1.0	68.66	76.78
1.2	67.02	74.75
1.5	62.63	69.51
2.0	52.95	58.19

Table 17: **Fixed-alpha Ada-HiSS-Merge transfer across task counts and pruning ratios.** All rows use $\alpha = 0.9$, global-threshold rank allocation, a top-1% sparse residual tail, and task-specific heads. Validation and held-out test averages are reported to show that the selected coefficient transfers beyond the validation setting used for selection.

Regime	Tasks	Val. top-1	Test top-1	Val. norm.	Test norm.
Full dense CLIP ViT-B/16	8	84.76	85.67	91.43	92.35
	14	83.25	83.27	89.74	90.47
	20	81.94	81.70	88.19	88.11
Pruned Taylor 0.3	8	83.45	83.28	90.69	90.57
	14	81.01	80.47	87.89	88.16
	20	80.27	80.00	86.85	86.84
Pruned Taylor 0.5	8	79.32	79.43	87.24	87.19
	14	77.12	76.65	84.27	84.74
	20	76.86	76.21	83.93	83.53
Pruned Taylor 0.7	8	69.87	70.42	77.48	78.22
	14	68.44	68.31	75.89	76.67
	20	68.74	67.92	76.94	76.09

Table 18: **Task-Update Compression component ablations on structured-pruned timm ViT-B/16.** The 8-task suite with full SUN397 evaluation is used. The spectral core is essential: sparse tails alone collapse. Global singular-value thresholding is better than fixed per-layer rank, and residual tails are slightly better than raw-update tails.

Ablation	k	p	Rank rule	Tail source	EuroSAT	DTD	Cars	SUN397	SVHN	RESISC45	MNIST	GTSRB	Avg.	Avg. norm
Spectral core only	0.25	0.00	global	residual	84.52	57.93	67.73	**39.89**	92.23	82.68	98.78	93.54	77.16	**88.56**
Ada-HiSS default	0.25	0.05	global	residual	84.85	57.98	67.50	39.63	92.14	**82.89**	98.80	**93.56**	77.17	88.53
Ada-HiSS best p	0.25	0.08	global	residual	85.26	58.19	67.37	39.48	92.19	**82.89**	**98.81**	93.44	**77.20**	88.55
Fixed per-layer rank	0.25	0.05	fixed	residual	84.44	58.09	65.71	38.01	92.22	82.57	**98.81**	93.47	76.66	87.77
Raw-update tail	0.25	0.05	global	raw update	**85.33**	**58.09**	67.32	39.52	**92.27**	82.27	98.76	93.42	77.12	88.47
Sparse tail only	0.25	0.05	none	raw update	24.15	19.79	4.68	9.89	15.84	13.19	48.56	5.87	17.75	20.25

Table 19: **Sensitivity to sparse residual budget p under Task-Update Compression.** These quick runs use the same 8-task structured-pruned timm ViT-B/16 setup with SUN397 capped to 2,000 examples, so they should be interpreted as a sensitivity diagnostic rather than the main benchmark protocol. Accuracy is stable over a broad low-budget range.

k	p	EuroSAT	DTD	Cars	SUN397 cap	SVHN	RESISC45	MNIST	GTSRB	Avg.	Avg. norm
0.25	0.00	84.52	57.93	67.73	33.95	92.23	82.68	98.78	93.54	76.42	87.11
0.25	0.01	84.52	58.03	67.65	33.65	92.23	82.57	98.79	93.59	76.38	87.04
0.25	0.02	84.89	58.14	67.53	33.90	92.24	82.68	98.78	93.62	76.47	87.16
0.25	0.03	84.74	57.98	**67.80**	33.90	92.21	82.78	98.78	**93.63**	76.48	87.17
0.25	0.05	84.85	57.98	67.50	**34.25**	92.14	**82.89**	98.80	93.56	76.50	87.22
0.25	0.08	**85.26**	58.19	67.37	34.10	92.19	**82.89**	**98.81**	93.44	**76.53**	**87.24**
0.25	0.10	85.04	**58.56**	67.27	33.55	92.14	82.81	98.79	93.51	76.46	87.11

G SUPPLEMENTARY DEIT EXPERIMENTS

This section provides the full tuned isomorphic-pruned DeiT experimental track summarized in Table 5. Table 20 reports the 8-, 14-, and 20-task results in a compact capacity-by-method format, and Table 21 reports the corresponding model and serialized-storage accounting. The evaluated capacities are `deit_4.2G`, `deit_2.6G`, `deit_1.2G`, and `deit_0.6G` from the Isomorphic-Pruning release. All downstream task experts use the same task lists as the main experiments and are fine-tuned with learning rate 5×10^{-5}, weight decay 0.01, and label smoothing 0.1. SUN397Val evaluation is capped at 2000 samples for runtime consistency. Ada-HiSS and TSVC use comparable task-update compression budgets with $k = 0.25$; Ada-HiSS reserves $p = 0.05$ for the sparse residual tail and uses per-task compact updates. This sparse budget is a residual-tail budget, not an evaluation merge scale. TIES uses its standard trimming budget, and EMR uses its shared-core/tail configuration.

H DENSE VERSUS PRUNED TIES DIAGNOSTIC

Table 22: **Small-Room diagnostic on the structured-pruned timm ViT-B/16 8-task suite.** Individual pruned experts retain high single-task accuracy, but TIES merging collapses more severely in the pruned regime. Recovery is merged accuracy divided by individual expert accuracy.

	Base accuracy		Merged accuracy (TIES)		Recovery	
Dataset	Dense	Pruned	Dense	Pruned	Dense	Pruned
MNIST	99.56	99.10	60.13	30.89	60.4	31.2
EuroSAT	99.22	98.78	43.22	16.63	43.6	16.8
RESISC45	96.24	94.60	40.38	16.19	41.9	17.1
DTD	82.29	74.26	35.11	18.99	42.7	25.6
SVHN	96.43	95.69	28.50	13.83	29.6	14.5
GTSRB	96.48	96.79	25.19	7.52	26.1	7.8
SUN397	72.68	51.24	13.70	8.07	18.9	15.8
Cars	85.44	77.89	10.22	2.85	11.9	3.7
Average	91.04	86.04	32.06	14.37	34.4	16.6

I PER-DATASET RESULTS

Table 23: **Per-dataset top-1 accuracy (%) on the structured-pruned timm ViT-B/16 8-task benchmark.** The TIES row uses the pruned TIES setting, matching the main 8-task average in Table 3. Bold entries mark the best method for each dataset.

Method	EuroSAT	DTD	Cars	SUN397	SVHN	RESISC45	MNIST	GTSRB	Avg.
TIES	16.63	18.99	2.85	8.07	13.83	16.19	30.89	7.52	14.37
EMR	65.70	53.19	29.09	25.79	72.63	75.71	96.44	85.23	62.97
TSVC	**84.89**	57.93	58.97	34.37	91.98	81.57	98.78	93.00	75.19
Ada-HiSS	84.85	**57.98**	**67.50**	**39.63**	**92.14**	**82.89**	**98.80**	**93.56**	**77.17**

Table 24: **Per-dataset top-1 accuracy (%) on the structured-pruned timm ViT-B/16 14-task benchmark.** Bold entries mark the best method for each dataset.

Method	Euro	DTD	Cars	FER	PCAM	C100	Flwr	OxPet	STL	SUN	SVHN	RES	MNIST	GTS	Avg.
TIES	10.22	10.37	1.42	18.00	54.10	5.57	2.55	26.76	50.33	5.04	13.74	5.71	16.92	3.25	15.99
EMR	64.15	39.10	13.61	51.10	85.39	45.47	42.35	81.49	91.73	18.36	62.96	63.37	95.10	72.11	59.02
TSVC	84.48	**58.24**	63.85	60.23	91.05	59.95	79.41	**88.20**	90.83	36.78	92.14	82.44	**98.80**	93.44	77.13
Ada-HiSS	**84.70**	58.09	**67.63**	**60.25**	**91.12**	**61.37**	**80.86**	87.98	**91.10**	**39.67**	**92.18**	**82.83**	98.78	**93.70**	**77.88**

Table 20: **Full DeiT comparison across pruning capacities.** Average top-1 and expert-normalized top-1 accuracy are reported for the tuned isomorphic-pruned DeiT family. Bold entries mark the best method within each model, task count, and metric.

		8 tasks		14 tasks		20 tasks	
Model	Method	Avg.	Norm.	Avg.	Norm.	Avg.	Norm.
deit_4.2G	TIES	33.35	39.49	32.77	37.98	32.11	38.28
	EMR	75.88	86.81	73.87	84.01	73.87	84.49
	TSVC	84.53	97.23	85.71	98.31	85.78	98.76
	Ada-HiSS	**84.86**	**97.74**	**85.93**	**98.63**	**85.97**	**99.02**
deit_2.6G	TIES	25.11	28.66	24.36	27.42	23.46	28.19
	EMR	71.94	80.09	68.89	76.79	69.10	77.95
	TSVC	85.15	97.30	86.01	98.31	85.95	98.73
	Ada-HiSS	**85.56**	**97.95**	**86.31**	**98.75**	**86.23**	**99.15**
deit_1.2G	TIES	24.08	30.68	21.54	25.61	20.70	26.42
	EMR	65.09	78.86	58.35	69.23	56.09	67.03
	TSVC	78.21	96.25	80.23	97.60	80.81	98.18
	Ada-HiSS	**78.34**	**96.47**	**80.44**	**97.93**	**81.00**	**98.50**
deit_0.6G	TIES	15.94	18.93	17.39	20.31	16.41	20.80
	EMR	60.50	71.10	56.81	67.06	56.71	66.90
	TSVC	76.92	94.23	78.79	96.32	79.71	97.18
	Ada-HiSS	**77.30**	**94.94**	**79.04**	**96.77**	**79.96**	**97.61**

Table 21: **Supplementary model and serialized-storage sizes.** DeiT source artifacts are counted with the ImageNet head; representative tuned expert artifacts use CarsVal task heads. TIMM ViT-B/16 rows give representative dense and structured-pruned task-expert sizes for comparison.

Family	Model	Artifact kind	Params	Reduction	FP32 storage	Serialized size
DeiT	DeiT-B/16 dense	architecture reference	86.568M	0.00%	330.23 MiB	–
DeiT	deit_4.2G	source, ImageNet head	20.694M	76.09%	78.94 MiB	237.05 MiB
DeiT	deit_2.6G	source, ImageNet head	13.069M	84.90%	49.86 MiB	149.80 MiB
DeiT	deit_1.2G	source, ImageNet head	5.738M	93.37%	21.89 MiB	65.90 MiB
DeiT	deit_0.6G	source, ImageNet head	3.076M	96.45%	11.74 MiB	35.44 MiB
DeiT	deit_4.2G	CarsVal tuned expert	20.075M	76.81%	76.58 MiB	76.68 MiB
DeiT	deit_2.6G	CarsVal tuned expert	12.576M	85.47%	47.97 MiB	48.08 MiB
DeiT	deit_1.2G	CarsVal tuned expert	5.428M	93.73%	20.71 MiB	20.81 MiB
DeiT	deit_0.6G	CarsVal tuned expert	2.843M	96.72%	10.85 MiB	10.95 MiB
TIMM ViT-B/16	dense	CarsVal tuned expert	85.949M	0.00%	327.87 MiB	327.98 MiB
TIMM ViT-B/16	pruned	CarsVal tuned expert	21.741M	74.70%	82.94 MiB	83.04 MiB
TIMM ViT-B/16	dense	SUN397Val tuned expert	86.104M	0.00%	328.46 MiB	328.57 MiB
TIMM ViT-B/16	pruned	SUN397Val tuned expert	21.819M	74.66%	83.23 MiB	83.33 MiB

Table 25: **Per-dataset top-1 accuracy (%) on the structured-pruned timm ViT-B/16 20-task benchmark, part I.** Bold entries mark the best method for each dataset.

Method	EuroSAT	DTD	Cars	KMNIST	EMNIST	SST2	FMNIST	Food101	CIFAR10	FER2013
TIES	13.26	7.07	0.90	13.42	12.25	49.92	10.87	3.93	21.72	14.54
EMR	74.11	27.34	2.96	69.24	93.14	49.92	79.79	9.56	79.33	35.82
TSVC	84.48	**58.24**	63.85	94.41	98.79	**50.41**	89.98	50.08	90.12	**60.23**
Ada-HiSS	**84.67**	58.14	**67.81**	**94.46**	**98.80**	**50.41**	**89.99**	**51.07**	**90.41**	60.03

Table 26: **Per-dataset top-1 accuracy (%) on the structured-pruned timm ViT-B/16 20-task benchmark, part II.** Bold entries mark the best method for each dataset.

Method	PCAM	C100	Flowers	OxPet	STL10	SUN397	SVHN	RESISC	MNIST	GTSRB	Avg.
TIES	60.11	2.20	4.62	15.94	35.11	2.19	12.68	4.65	18.47	3.78	15.38
EMR	85.64	16.76	23.55	60.37	90.28	5.49	57.15	25.24	90.79	53.53	51.50
TSVC	91.05	59.95	79.41	**88.20**	90.83	36.78	92.14	82.44	**98.80**	93.44	77.68
Ada-HiSS	**91.11**	**61.56**	**81.07**	88.09	**91.03**	**39.84**	**92.23**	**82.71**	98.78	**93.59**	**78.29**