A CONTROLLED BENCHMARK FOR WEIGHT-SPACE MODEL MERGING ACROSS MEDICAL IMAGING DOMAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Consolidating medical-imaging specialists into a shared model requires preserving performance across distinct imaging domains. Weight-space model merging offers a way to combine independently fine-tuned encoders without retraining, but comparisons between merging methods are sensitive to validation protocols. Official evaluation scripts for TSV-Merge and isotropic merging use wider scaling-coefficient ranges for selected spectral methods than for other tunable baselines. We introduce a controlled benchmark spanning dermoscopy, chest radiography, and histopathology, with three binary tasks and $2{,}000$ training images per specialist. We compare 11 methods across seven vision-transformer backbones and three seeds using a shared scaling-coefficient grid. The best observed merger reaches 0.811 mean AUROC against 0.853 for task-aware specialist routing, with tuned Task Arithmetic close behind at 0.808. We also introduce Gram-LS, a closed-form method based on task-vector projection constraints that performs comparably to tuned summation. Within-method validation tuning improves Gram-LS and Task Arithmetic by approximately 0.030 to 0.035 AUROC, exceeding the 0.019 spread among five leading tuned methods, while search bounds can change method rankings. Iso-C and Iso-CTS fall 0.10 to 0.13 AUROC below tuned summation. These findings establish tuned summation as a strong baseline in this setting and show that specialist references and validation protocols are essential to assessing the performance cost of consolidation.

1 INTRODUCTION

Weight-space *model merging* combines specialists fine-tuned from a common pretrained model without retraining or accessing their training data (Ilharco et al., 2023; Yadav et al., 2023). In medical imaging, this approach could consolidate encoders specialized for different imaging modalities while retaining task-specific prediction heads. Consolidation reduces parameter storage, but changes the representation used by each head and may therefore reduce task performance. Evaluating this tradeoff requires measuring both the performance retained relative to individual specialists and the sensitivity of the result to the merging and validation procedures.

MergeBench standardizes comparisons of domain-specialized language models (He et al., 2025), and FusionBench supports evaluation across model-fusion settings (Tang et al., 2025). Yet concrete protocol differences remain in widely used vision evaluations. The official scripts accompanying TSV-Merge and isotropic merging search scaling coefficients over $(0, 3]$ for selected spectral methods, while other tunable baselines use $(0, 1]$ (Gargiulo et al., 2025; Marczak et al., 2025).[1] This asymmetry can conflate the effect of the merging rule with the range of configurations considered. Medical tasks additionally differ in task formulation, class count, and dataset size. A controlled comparison must account for these differences and measure performance relative to the specialists being consolidated.

We address these issues with a controlled benchmark spanning dermoscopy, chest radiography, and histopathology. The benchmark standardizes three tasks to binary detection under AUROC and limits

[1]The default search branches in `evaluate_task_vector`, in `src/eval/eval.py`: TSV-Merge and isotropic merging. Fixed-coefficient methods and explicit coefficient overrides use separate branches.

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.

each specialist to 2,000 training images. We compare 11 merging methods across seven vision-transformer backbones and three seeds. Merged models retain task-specific single-logit heads, and both merged models and the specialist reference use the known task identity at inference. Methods with tunable scaling share a common scaling-coefficient grid, and we report the method-specific search spaces explicitly.

The best observed merger reaches 0.811 mean AUROC against 0.853 for specialist routing. Tuned Task Arithmetic follows at 0.808, and the five leading methods span 0.019 AUROC. Within-method tuning improves Task Arithmetic by 0.035, exceeding this spread, and expanding the scaling-coefficient range raises LiNeS from 0.763 to 0.811. Iso-C and Iso-CTS reach 0.681 and 0.704.

To investigate the role of task-vector interactions, we introduce *Gram-LS*, a closed-form method that uses the task-vector Gram matrix to preserve projections onto individual task directions. Gram-LS reaches 0.808 mean AUROC, matching tuned summation without outperforming it, and within-method tuning improves it by 0.030, in line with the gain for Task Arithmetic. For orthogonal task vectors, its unregularized solution reduces exactly to summation. We complement this comparison with a synthetic analysis showing how isotropization attenuates dominant task directions, and an exploratory ablation associating Gram-matrix conditioning with the magnitude of differences between Gram-LS and TSV-Merge.

Summary of contributions

- **A controlled medical-imaging benchmark.** We compare 11 merging methods across seven ViT backbones and three seeds, standardizing task formulation and training-image count and evaluating against task-aware specialist routing (§3.1; Table 1).
- **Evidence that tuning protocols materially affect merger comparisons.** Within-method tuning gains exceed the observed spread among five leading methods, and expanding the scaling range changes LiNeS's ranking. Under the shared scaling grid, tuned summation remains competitive with more elaborate methods (§4; Appendix B).
- **A projection-based method and geometric analysis.** We introduce Gram-LS to test whether accounting for task-vector interactions improves merging, and examine how projection constraints, singular spectra, and Gram-matrix conditioning relate to merger performance (§3.3; §4.1).

2 Related work

Training-free mergers combine task vectors $\tau_i = \theta_{\mathrm{ft},i} - \theta_{\mathrm{pre}}$, differing primarily in parameter weighting. Uniform weight averaging and Task Arithmetic (Ilharco et al., 2023) scale vectors using a single shared scalar. To reduce inter-task interference, coordinate-wise methods modify internal parameters independently. TIES trims small entries and enforces sign consensus (Yadav et al., 2023), DARE randomly drops and rescales task-vector entries (Yu et al., 2024), PCB-Merging balances parameter competition (Du et al., 2024), and LiNeS scales weights by layer depth (Wang et al., 2025).

Fisher-weighted merging (Matena & Raffel, 2022) weights parameters using estimated Fisher information, while RegMean (Jin et al., 2023) minimizes a regression objective using activation Gram matrices $X^\top X$. MaTS (Tam et al., 2024) generalizes merging to linear systems defined by task parameter subspaces, including activation- and Fisher-based formulations, and solves them with conjugate gradients. Spectral methods operate directly on layer updates: TSV-Merge (Gargiulo et al., 2025) orthogonalizes retained singular directions, while Iso-C and Iso-CTS (Marczak et al., 2025) flatten singular spectra in common and task-specific subspaces. These methods use singular directions to represent and modify relationships among task updates.

Gram-LS is related to closed-form methods based on task-vector geometry. MetaGPT (Zhou et al., 2024) derives task-vector norm weights, while NAN (Si et al., 2025) uses inverse model-weight norms. A data-free variant of Bayesian Model Merging and ACE-Merging estimate input second-moment statistics from model weights (Li et al., 2026; Xu et al., 2026). Gram-LS instead imposes projection constraints directly on the merged weight update. The resulting linear system uses the $K \times K$ matrix of task-vector inner products and reduces to summation for mutually orthogonal, nonzero task vectors.

This formulation provides a direct test of whether adjusting for these inner products improves on a scaled sum.

Prior merging studies already compare merged models with individual specialists, including RegMean and Iso-C/CTS (Jin et al., 2023; Marczak et al., 2025). TSV-Merge distinguishes specialist compression when task identity is available (TSV-C) from merging when it is not (TSV-M) (Gargiulo et al., 2025). PHATGOOSE evaluates expert routing against a merged-expert baseline (Muqeeth et al., 2024). We use task-aware specialist routing as the reference for quantifying the performance cost of consolidation.

MergeBench and FusionBench support systematic evaluation across merging settings (He et al., 2025; Tang et al., 2025), and Yadav et al. (2025) find that differences between merging methods diminish as model size increases. In medical imaging, MedMerge learns kernel-level merging weights for transfer learning (Almakky et al., 2024), while Ambekar et al. (2026) study test-time entropy-adaptive merging with separate encoder and classification-head coefficients. We evaluate static, training-free mergers across seven backbones under a common medical task protocol.

3 METHODS

3.1 A CONTROLLED BINARY BENCHMARK

We fine-tune 7 pretrained ViT backbones (CLIP ViT-B/16, ViT-B/16, DINOv3 ViT-S/16, RAD-DINO, DINOv2-Base, ViT-MAE-Base, BEiT-Base) into per-dataset specialists on a controlled trio of binary tasks spanning three disjoint medical-imaging domains, namely melanoma detection (ISIC 2017 dermoscopy), pleural-effusion detection (CheXpert chest radiography), and tumor detection (PatchCamelyon lymph-node histopathology). Fixing a single task type (binary), a single evaluation metric (AUROC), and task-specific single-logit heads on a shared encoder removes the task-type, class-count, and metric confounds that otherwise entangle cross-domain merging comparisons, and makes the evaluation protocol comparable across tasks. Low-rank update structure is an empirical property rather than a consequence of binary labels alone. To also control for dataset scale, we train each specialist on exactly $2{,}000$ images, drawn as a seed-deterministic subsample of the task's training pool. The three pools differ by two orders of magnitude, from $2{,}200$ images for ISIC 2017 to $262{,}144$ for PatchCamelyon (Appendix A), so without this cap dataset size would vary with domain. Specialists use AdamW with cosine warmup and early stopping at LR $= 5\mathrm{e}{-5}$, with CLIP at $1\mathrm{e}{-5}$. We repeat each configuration over 3 seeds.

3.2 MERGING METHODS

We merge task vectors $\tau_i = \theta_{\mathrm{ft},i} - \theta_{\mathrm{pre}}$, computed over the shared ViT encoder parameters, across 11 methods, namely Simple Averaging, Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023), DARE (Yu et al., 2024), DARE-TIES, PCB-Merging (Du et al., 2024), LiNeS (Wang et al., 2025), Iso-C with its stronger Iso-CTS variant (Marczak et al., 2025), TSV-Merge (Gargiulo et al., 2025), and Gram-LS. We select each method's hyperparameters by grid search on the validation split, using one scaling-coefficient grid for methods with tunable scaling, which we widened until boundary selection nearly vanished (Appendix B). Heads remain task-specific. At inference, the known dataset identity selects the corresponding specialist head on the shared merged encoder. Thus both the merged model and specialist reference assume access to task identity.

3.3 GRAM-LS MERGE

Gram-LS incorporates pairwise task-vector inner products through a closed-form system of projection constraints. For mutually orthogonal task vectors, uniform averaging, $m = \frac{1}{K}\sum_i \tau_i$, retains $1/K$ of each task's original projection, regardless of its norm. Summation restores these projections in the orthogonal case. For general task vectors, cross-task inner products also contribute, motivating coefficients that satisfy the projection constraints jointly. We solve for task coefficients $c \in \mathbb{R}^K$ in $m = \sum_j c_j \tau_j$ such that the projection of m onto each task direction recovers that task's original magnitude, so that $\langle m, \tau_i \rangle = \langle \tau_i, \tau_i \rangle$ for all i. This constraint yields the linear system

$$G\,c = \mathrm{diag}(G), \qquad G_{ij} = \langle \tau_i, \tau_j \rangle, \tag{1}$$

Table 1: Mean test AUROC on the controlled binary trio across 7 backbones $\times$ 3 seeds ($n = 21$ cells). Each entry carries the half-width of a 95% confidence interval on the cell mean, $t_{0.975,n-1}\, s/\sqrt{n}$, computed per task and for the three-task mean. **Bold** denotes methods not statistically separable from the best via paired t-test over shared cells. Because the intervals are marginal, they absorb between-backbone heterogeneity and are therefore much wider than the paired differences the bolding is based on (§4). The router uses individual specialists as a reference.

Method	Melanoma	Effusion	Tumor	Mean
Individual specialists (routing)	*0.804* ±0.028	*0.807* ±0.020	*0.949* ±0.007	*0.853* ±0.011
Simple Averaging	0.717 ±0.025	0.697 ±0.039	0.880 ±0.030	0.765 ±0.027
Task Arithmetic	0.752 ±0.034	0.761 ±0.023	0.910 ±0.021	**0.808** ±0.021
TIES	0.737 ±0.038	0.742 ±0.038	0.909 ±0.025	0.796 ±0.027
DARE	0.753 ±0.034	0.762 ±0.022	0.911 ±0.021	**0.808** ±0.021
DARE-TIES	0.743 ±0.040	0.748 ±0.033	0.915 ±0.016	0.802 ±0.025
PCB-Merging	0.757 ±0.026	0.720 ±0.052	0.899 ±0.036	**0.792** ±0.035
LiNeS	0.755 ±0.034	0.761 ±0.021	0.918 ±0.014	**0.811** ±0.019
Iso-C (Marczak et al., 2025)	0.632 ±0.028	0.625 ±0.028	0.786 ±0.044	0.681 ±0.025
Iso-CTS (Marczak et al., 2025)	0.658 ±0.028	0.648 ±0.030	0.807 ±0.041	0.704 ±0.025
TSV-Merge (Gargiulo et al., 2025)	0.751 ±0.036	0.744 ±0.032	0.910 ±0.017	0.801 ±0.024
Gram-LS (ours)	0.749 ±0.039	0.755 ±0.031	0.919 ±0.014	**0.808** ±0.023

with the closed-form solution $c = G^{-1}\,\mathrm{diag}(G)$ when G is nonsingular. We compute c independently per weight tensor to adapt to layer-specific geometry, and we stabilize the solve with a relative Tikhonov ridge $G + \lambda\bar{g}I$ (where $\bar{g}$ is the mean diagonal of G).

For mutually orthogonal, nonzero task vectors, G is diagonal and the unregularized solution is $c = \mathbf{1}$, so Gram-LS reduces to summation. For correlated task vectors, the coefficients depend on both the diagonal and off-diagonal entries of G. These constraints concern weight-space projections and do not guarantee preservation of specialist predictions. The merged model weights are $\theta_{\mathrm{pre}} + \alpha \sum_j c_j \tau_j$. For nonsingular G, setting $\alpha = 1$ and $\lambda = 0$ satisfies the projection constraints exactly, and ridge regularization relaxes them. We evaluate this validation-free default alongside a parameterization in which α and λ are selected on validation data. The weight-combination step requires no training or data access, whereas hyperparameter selection requires validation data.

3.4 Evaluation

We report per-task AUROC across every (backbone, method, seed) experimental cell. We quantify the performance penalty of merging as $\Delta_d = \mathrm{oracle}_d - \mathrm{merged}_d$, where the oracle router directs each test sample to its dataset-specific specialist, providing a reference for retaining specialists, not a theoretical upper bound on merged performance. Because we evaluate every method on identical configurations, we compare methods by paired within-cell differences rather than by marginal means alone, so that high inter-backbone variance does not mask method-level distinctions.

4 Experiments and results

Table 1 reports mean AUROC across 7 backbones and 3 seeds. The methods separate into a tight leading cluster within 0.019 of one another, a middle group, and a weaker isotropic family, all below the specialist reference.

Comparative analysis of merging methods. LiNeS achieves the highest observed mean (0.811 AUROC), followed by Gram-LS, Task Arithmetic, and DARE at 0.808 AUROC. PCB-Merging reaches 0.792 AUROC with higher variability. None of the ten unadjusted pairwise comparisons among these five methods reaches the 5% significance threshold (maximum $t = 2.00$). The estimates nevertheless span 0.019 AUROC, leaving potentially meaningful performance differences unresolved, particularly between PCB-Merging and LiNeS. We report the secondary comparisons favoring LiNeS over TSV-Merge ($+0.010$, $t = 2.42$), DARE-TIES ($+0.009$, $t = 2.54$), and TIES ($+0.015$, $t = 3.10$) as exploratory, since we do not correct them for multiple testing. Gram-LS performs comparably

Table 2: Memory cost of routing to three specialists versus serving one merged model.

Backbone	Encoder	3 specialists	Merged	Saved (fp32)
CLIP ViT-B/16	85.8M	257.4M	85.8M	686 MB
ViT-B/16	86.4M	259.2M	86.4M	691 MB
DINOv3 ViT-S/16	21.6M	64.8M	21.6M	173 MB
RAD-DINO	86.6M	259.7M	86.6M	693 MB
DINOv2-Base	86.6M	259.7M	86.6M	693 MB
ViT-MAE-Base	85.8M	257.4M	85.8M	686 MB
BEiT-Base	85.8M	257.3M	85.8M	686 MB

to tuned scaled summation without demonstrating a distinct advantage. Against the specialist baseline of 0.853 AUROC, LiNeS gives up 0.042 AUROC, and Gram-LS and Task Arithmetic give up approximately 0.045 AUROC. These gaps are measured at the evaluated training-image budget.

Marginal intervals reflect backbone heterogeneity rather than seed noise. The per-task intervals in Table 1 are wide because they pool two distinct sources of variation. Averaged over the eleven evaluated methods, the standard deviation across the three seeds within a fixed backbone is 0.035 AUROC on melanoma, 0.035 on effusion, and 0.030 on tumor, whereas the standard deviation of the seven per-backbone means is 0.067, 0.059, and 0.040 respectively. Backbone identity thus contributes roughly twice the spread of seed resampling. A marginal interval computed over all 21 cells consequently overstates the uncertainty of a method-versus-method comparison, because both methods are evaluated on the same backbones and inherit the same backbone effect. Paired tests over shared cells remove that common component, so differences of 0.010 AUROC can separate methods whose marginal intervals overlap almost entirely. Appendix E reports every method at every backbone with its across-seed standard deviation.

Memory savings. By consolidating N task specialists into a single encoder backbone, model merging eliminates redundant parameter storage, saving up to 693 MB per deployment for ViT-B models (Table 2), while task-specific heads introduce negligible parameter overhead.

Impact of hyperparameter search spaces. The performance rankings reported in Table 1 depend heavily on the choice of validation grid boundaries. Restricting the scaling-coefficient search ceiling to 0.9 reduces the observed performance of LiNeS to 0.763 AUROC, ranking it near the bottom of all evaluated techniques. Expanding this upper bound to 4.0 elevates LiNeS to 0.811 AUROC, placing it in the top tier, while leaving methods with optima inside the original interval unaffected. Grid bounds alone can therefore reorder competitive leaderboards. This sensitivity illustrates why the unequal search ranges identified in the introduction matter for method comparisons.

Evaluation of validation-free strategies. Table 3 reports four validation-free strategies, namely unscaled Task Arithmetic ($\alpha = 1$), Gram-LS ($\alpha = 1$, $\lambda = 0$), and norm-based weighting adapted from MetaGPT (Zhou et al., 2024) and NAN (Si et al., 2025). Both adaptations weight task vectors; the NAN-style variant therefore differs from the original model-weight formulation. The norm-based variants perform similarly to uniform averaging. Untuned Gram-LS reaches 0.777 AUROC and unscaled Task Arithmetic reaches 0.773. Selecting hyperparameters on validation data raises both methods to 0.808. Computed from unrounded scores over 21 paired backbone-seed cells, the gains are 0.030 AUROC for Gram-LS ($t = 2.93$) and 0.035 for Task Arithmetic ($t = 3.19$). Gram-LS tunes scaling and ridge regularization, whereas Task Arithmetic tunes scaling alone. Both within-method gains exceed the 0.019 spread among the five leading tuned methods. The paired tests use the unadjusted cell-level analysis described in §3.4.

No detected transfer advantage over individual specialists. A key argument for model merging over routing is that a merged encoder might generalize to domains absent from the training suite, where no dedicated expert exists for routing. TIES and RegMean report gains on held-out tasks or domains (Yadav et al., 2023; Jin et al., 2023). Model soups improve out-of-distribution robustness by averaging models from a hyperparameter sweep on the same task (Wortsman et al., 2022). We examine transfer from merged medical specialists via linear probes on frozen encoders across two

Table 3: Performance of validation-free merging strategies across the 21 evaluation settings from Table 1, reported with 95% confidence intervals on cell means.

Method (zero tuning)	Mean AUROC
Raw sum ($\alpha = 1$)	0.773 $\pm$0.041
Gram-LS ($\alpha = 1, \lambda = 0$)	0.777 $\pm$0.041
MetaGPT-style norm weighting	0.764 $\pm$0.018
NAN-style inverse-norm weighting	0.765 $\pm$0.029
Simple averaging (Table 1)	0.765 $\pm$0.027

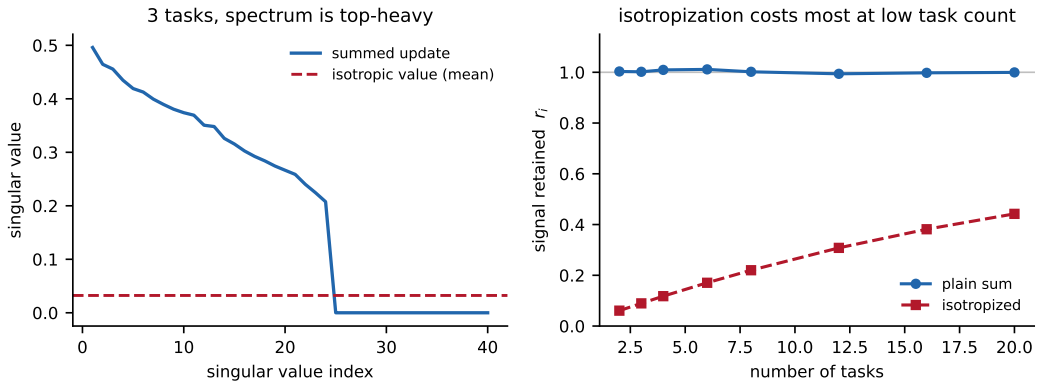


Figure 1: Effect of isotropization on synthetic low-rank updates. **Left**, the summed update of three tasks has a concentrated singular spectrum, with dominant singular values substantially above the isotropic value. **Right**, signal retention $r_i = \langle m, \tau_i \rangle / \langle \tau_i, \tau_i \rangle$. Summation maintains $r_i \approx 1$ across the tested task counts, while signal retention under isotropization increases with task count.

held-out medical datasets (21 cells), RetinaMNIST (a distant modality, fundus photography) and HAM10000 (a near modality, dermoscopy). On RetinaMNIST, the best merged encoder records 0.431 balanced accuracy against 0.437 for the top individual specialist and 0.446 for the un-tuned pretrained backbone. On HAM10000, accuracy reaches 0.601 for the merged model, 0.596 for the best specialist, and 0.566 for the base model. Task fine-tuning yields gains only when the target domain aligns with fine-tuned tasks, improving over the base model on HAM10000 ($+0.034$, $t = 2.52$) while producing no benefit on RetinaMNIST (-0.015, $t = -1.38$). Merging provides no statistical advantage over the single best specialist in either regime ($+0.005$ near with $t = 0.71$, and -0.006 far with $t = -0.75$).

Isotropic subspace methods underperform in the evaluated setting. TSV-Merge reaches 0.801 AUROC, whereas Iso-C and Iso-CTS reach 0.681 and 0.704. Synthetic low-rank updates illustrate a possible explanation, since replacing singular values with their mean can attenuate dominant task directions. In this synthetic setting, signal retention is below 10% at $N = 3$ tasks and reaches 44% at $N = 20$ (Figure 1). Iso-C also underperforms tuned summation in a six-task evaluation (0.686 versus 0.775 AUROC; Appendix C).

4.1 Task-vector geometry

To examine how task-vector geometry relates to merger performance, we conduct a separate exploratory comparison using ISIC 2017, CheXpert, and PathMNIST across seven backbones and three seeds, with a 2,000-image training cap. Both arms draw from these source datasets, and PathMNIST replaces the primary benchmark's PatchCamelyon. The binary arm uses melanoma, effusion, and tumor detection and averages AUROC. The native arm uses 3-class dermoscopy, 5-label radiology, and 9-class histopathology, averaging balanced accuracy for ISIC and PathMNIST with macro-AUROC for CheXpert. The launch and loader code use shared source pools and seed-deterministic

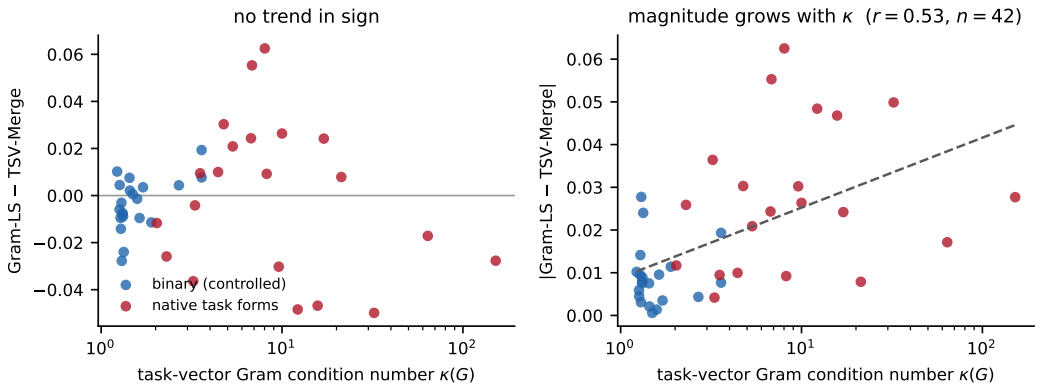


Figure 2: Conditioning is associated with the magnitude of relative performance differences in the exploratory ablation. Each point represents one (backbone, seed) cell. Binary gaps use mean AUROC, and native gaps mix balanced accuracy and macro-AUROC. **Left:** The Gram-LS minus TSV-Merge gap fans out with condition number $\kappa(G)$ without displaying a monotonic trend. **Right:** The absolute gap magnitude grows with $\kappa(G)$ ($r = 0.53$, $n = 42$).

subsampling, but the comparison changes evaluation metrics as well as label structure. It therefore describes two task protocols rather than isolating a causal effect of label structure.

Averaged over cells, the reported tests do not detect a mean difference between Gram-LS and TSV-Merge in either arm, by -0.003 under the binary formulation and -0.001 under the native one, with $n = 21$ cells each. The spread changes instead, with the standard deviation of the Gram-LS minus TSV-Merge gap growing from 0.011 to 0.033, a descriptive variance ratio of 8.6 across the two metric conventions, and the native arm ranges from -0.050 to $+0.063$ where the binary arm stays inside ± 0.03. Under the native formulation Gram-LS is sometimes clearly better and sometimes clearly worse, and which one is not predictable from the domain.

The median per-layer condition number of the task-vector Gram matrix increases from $\kappa = 1.3$ under the binary formulation to $\kappa = 8.0$ under the native one, with individual native cells reaching $\kappa > 150$. Across all 42 cells (Figure 2), $\log \kappa$ predicts the magnitude of the Gram-LS minus TSV-Merge gap ($r = 0.531$, $t = 3.96$, $p < 0.001$) but not its sign ($r = -0.08$, n.s.). One possible explanation is numerical sensitivity. Solving $Gc = \mathrm{diag}(G)$ inverts G, so an ill-conditioned G can amplify perturbations in the task vectors. RegMean (Jin et al., 2023) also reports degenerate results with unmodified activation Gram matrices and attenuates their off-diagonal entries. This operation differs from the diagonal ridge used in Gram-LS. A related geometric account of task arithmetic links successful task composition to weight disentanglement (Ortiz-Jimenez et al., 2023).

The pooled correlation may reflect differences between ablation arms or backbones. Within-arm, backbone-aware estimates are needed to distinguish these effects before using conditioning to select a merger. For an exactly diagonal, nonsingular G, the unregularized solution is $c = \mathbf{1}$, so Gram-LS reduces to summation. This identity helps explain the similar behavior of Gram-LS and summation when task vectors are nearly orthogonal. The primary benchmark reports 0.808 for both Gram-LS and Task Arithmetic to three decimal places. The separate binary ablation reports a mean Gram-LS advantage of approximately 0.004 and should not be conflated with that benchmark.

Additional exploratory measurements report median per-layer pairwise cosine similarities no greater than 0.14 across the tested task-relatedness settings, and whole-vector Gram-LS coefficients c_j between 0.90 and 0.98 for six of seven backbones (DINOv2 is the exception).

Task-vector norms. Differences in task-vector norms can cause some updates to dominate averaging (Yoshida et al., 2026). Under our controlled binary protocol, task-vector norms remain uniform across distinct medical domains (domain means 6.4, 6.1, and 6.7), although domain-averaged norms alone do not isolate the effect of domain semantics. The observed norm range is substantially larger across backbones, with per-task norms ranging from 0.6 to 0.8 for CLIP to ≈ 22 for ViT-B. CLIP's smaller norms coincide with its $5\times$ lower fine-tuning learning rate, suggesting an optimization

contribution that cannot be separated from backbone effects without controlled learning-rate and step-count ablations.

In our native-task ablation, the 5-label multilabel classification task (CheXpert) produced the largest-norm updates regardless of dataset size, which is a task-type confound that naive comparisons overlook. Matching image counts and task formulation reduces selected confounds, but early stopping means that equal dataset sizes do not necessarily imply equal optimizer-step or compute budgets.

5 Discussion and limitations

The benchmark identifies two distinct performance gaps. Consolidating specialists into one encoder costs 0.042 AUROC for the best observed merger, while within-method validation tuning improves Gram-LS by 0.030 and Task Arithmetic by 0.035. These tuning gains exceed the 0.019 spread among the five leading tuned methods. The LiNeS search-range experiment further shows that scaling bounds can change a method's position in the ranking.

Synthetic low-rank updates show that isotropization can attenuate dominant task directions, and the exploratory ablation associates Gram-matrix conditioning with variation in the Gram-LS versus TSV-Merge gap. The held-out transfer evaluations find no detected advantage over the best individual specialist, leaving reduced encoder storage as the demonstrated benefit of consolidation in these experiments.

Limitations.

- **Architectural and task scope:** We evaluate only Vision Transformer (ViT) backbones at 224×224 resolution on standardized binary classification tasks, in a static, one-shot merging setup under a fixed training-image budget.
- **Training-set size:** Every specialist sees exactly $2{,}000$ images, a cap we impose to remove dataset scale as a confound between domains. We do not measure how the routing gap or the method ranking moves as that budget grows. Changing the budget may affect both specialist performance and task-vector geometry, so the reported 0.042 AUROC gap is specific to this operating point.
- **Task scaling dynamics:** We evaluate task sets of $N = 3$ and $N = 6$, with the primary benchmark using $N = 3$. Consequently, we cannot fully characterize how matrix conditioning (κ) or isotropic subspace methods evolve across much larger task scales ($N \geq 20$).
- **Conditioning diagnostics:** The conditioning association is exploratory, pooled across ablation arms, and does not establish causation or a validated method-selection rule. The arms use different metric conventions, so their gap magnitudes are not a metric-controlled comparison.
- **Statistical inference:** Reported paired tests are unadjusted for multiple comparisons and treat backbone-seed cells as the analysis units. Generalization across backbones requires accounting for seeds nested within backbones. We do not establish practical equivalence between mergers.
- **Baseline validation:** We exclude Fisher-weighted merging because its binary-loss and Fisher-estimation implementation requires correction and a rerun.
- **Clinical safety and demographic equity:** We do not evaluate whether parameter merging disproportionately degrades predictive performance across demographic subgroups (e.g., age, sex, ethnicity).

Future work. Varying the per-specialist training budget below the ISIC-bounded ceiling of $2{,}000$ images would show whether the specialist gap and the method ordering move with budget. We also plan to replace the pooled conditioning correlation with within-arm, backbone-aware estimates and to test $\kappa(G)$ prospectively as a switch between a scaled sum and a low-rank merge. Larger task suites ($N \geq 20$) would show whether the isotropization deficit closes as the synthetic model predicts.

6 Conclusion

On a controlled benchmark of three medical-imaging tasks, the best observed merger reaches 0.811 mean AUROC against 0.853 for task-aware specialist routing, while reducing encoder storage by two

thirds. Tuned Task Arithmetic and Gram-LS both reach 0.808. Within-method tuning gains of 0.030 and 0.035 AUROC exceed the 0.019 spread among five leading tuned methods, and widening the scaling range raises LiNeS from 0.763 to 0.811. Iso-C and Iso-CTS underperform tuned summation in this setting. These results establish tuned summation as a strong baseline for the evaluated tasks and show that consolidation costs and validation protocols are central to interpreting merger performance.

AI USE STATEMENT

We used Anthropic's Claude through Claude Code throughout this work, and OpenAI's Codex to review and revise the manuscript, inspect research code, and check reported numerical consistency. We used it to implement methods (the merging methods, data loaders, training and evaluation pipelines, and cluster scripts), to provide feedback on experimental design (the confidence-interval reporting and the layout of the ablations). We did not use generative AI to generate datasets, to develop the theoretical framing, to formulate or prove mathematical claims, to propose the hypotheses, to translate, to clean or reformat datasets, or for qualitative analysis. We used it to create and edit software code, to create figures through plotting scripts, and to format references. The authors reviewed and tested all AI-generated code, verified every reported number from the per-cell metric files with scripts they inspected, and revised all written text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The main benchmark and reported auxiliary analyses use public medical imaging datasets, including ISIC 2017, CheXpert, PatchCamelyon, PathMNIST, RetinaMNIST, and HAM10000 under their respective licenses and intended research uses. The dataset providers de-identified all images before release.

REPRODUCIBILITY STATEMENT

We specify the controlled benchmark in §3.1, with per-dataset splits, sizes, and class balances in Appendix A and the complete per-method hyperparameter grids and trial counts in Appendix B. We define Gram-LS in full in §3.3, where it is deterministic given the task vectors and requires no data at merge time. We describe our evaluation protocol, including the paired within-cell comparison behind every significance claim, in §3.4.

We checked dataset keys and metric definitions in all 126 result files and reproduced the 42 per-cell merger differences in the saved summary. The supplementary archive `anonymous-supplement.zip` contains anonymized source code, configurations, launch scripts, per-cell merger results, held-out probe results, and scripts for both figures. A dependency-free verification script recomputes from the packaged records every table and every reported statistic that does not require checkpoints, namely Table 1 with its confidence intervals and the specialist-routing baseline, Table 3 and the within-method tuning gains, the paired comparisons and spread decomposition of §4, the per-backbone tables of Appendix E, the six-task means of Appendix C, the 42 ablation cells with their gap and conditioning statistics, and the held-out probe comparison. The archive includes all 63 specialist test scores and excludes images and model checkpoints. The archive provides condition numbers and task-vector norms as saved measurements, and the verification script does not recompute them.

REFERENCES

Ibrahim Almakky, Santosh Sanjeev, Anees Ur Rehman Hashmi, Mohammad Areeb Qazi, Hu Wang, and Mohammad Yaqub. MedMerge: Merging Models for Effective Transfer Learning to Medical Imaging Tasks. *arXiv preprint arXiv:2403.11646*, 2024. URL `https://arxiv.org/abs/2403.11646v2`. Version 2, revised April 2025.

Sameer Ambekar, Reza Nasirigerdeh, Peter J. Schuffler, Lina Felsner, Daniel M. Lang, and Julia A. Schnabel. The Mean is the Mirage: Entropy-Adaptive Model Merging under Heterogeneous Domain Shifts in Medical Imaging. *arXiv preprint arXiv:2602.21372*, 2026. URL `https://arxiv.org/abs/2602.21372`.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547. URL https://arxiv.org/abs/1710.05006.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim Kuan Goh, Ho-Kin Tang, Daojing He, and Min Zhang. Parameter Competition Balancing for Model Merging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2410.02396.

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 (CVPR)*, 2025. URL https://arxiv.org/abs/2412.00081. arXiv:2412.00081.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. MergeBench: A Benchmark for Merging Domain-Specialized LLMs. In *NeurIPS Datasets and Benchmarks*, 2025. URL https://arxiv.org/abs/2505.10833. arXiv:2505.10833.

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 (ICLR)*, 2023. URL https://arxiv.org/abs/2212.04089.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://arxiv.org/abs/1901.07031.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless Knowledge Fusion by Merging Weights of Language Models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.09849. arXiv:2212.09849.

Kaiyang Li, Shaobo Han, Qing Su, and Shihao Ji. Bayesian Model Merging. *arXiv preprint arXiv:2605.12843*, 2026. URL https://arxiv.org/abs/2605.12843.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost van de Weijer. No Task Left Behind: Isotropic Model Merging with Common and Task-Specific Subspaces. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.04959. arXiv:2502.04959.

Michael Matena and Colin Raffel. Merging Models with Fisher-Weighted Averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2111.09832.

Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. Learning to Route Among Specialized Experts for Zero-Shot Generalization. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.05859. arXiv:2402.05859.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.12827. arXiv:2305.12827.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w. URL https://doi.org/10.1038/s42256-024-00965-w.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020.

Chongjie Si, Kangtao Lv, Jingjing Jiang, Yadao Wang, Yongwei Wang, Xiaokang Yang, Wenbo Su, Bo Zheng, and Wei Shen. NAN: A Training-Free Solution to Coefficient Estimation in Model Merging. *arXiv preprint arXiv:2505.16148*, 2025. URL https://arxiv.org/abs/2505.16148.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Derek Tam, Mohit Bansal, and Colin Raffel. Merging by Matching Models in Task Parameter Subspaces. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2312.04339. arXiv:2312.04339.

Anke Tang, Li Shen, Yong Luo, Enneng Yang, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. FusionBench: A Unified Library and Comprehensive Benchmark for Deep Model Fusion. *Journal of Machine Learning Research*, 26(307):1–38, 2025. URL https://www.jmlr.org/papers/v26/25-1243.html. arXiv:2406.03280.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018. doi: 10.1038/sdata.2018.161. URL https://www.nature.com/articles/sdata2018161.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation Equivariant CNNs for Digital Pathology. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018. URL https://arxiv.org/abs/1806.03962.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training Layer Scaling Prevents Forgetting and Enhances Model Merging. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.17146.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2203.05482.

Bo Xu, Haotian Wu, Hehai Lin, Weiquan Huang, Beier Zhu, Yao Shu, and Chengwei Qin. ACE-Merging: Data-Free Model Merging with Adaptive Covariance Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2603.02945. arXiv:2603.02945.

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 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.01708.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What Matters for Model Merging at Scale? *Transactions on Machine Learning Research (TMLR)*, 2025. URL https://openreview.net/forum?id=9sbetmvNpW. arXiv:2410.03617.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2—A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8. URL https://www.nature.com/articles/s41597-022-01721-8.

Kotaro Yoshida, Yuji Naraki, Takafumi Horie, Ryotaro Shimizu, Ioannis Mitliagkas, and Hiroki Naganuma. DisTaC: Conditioning Task Vectors via Distillation for Robust Model Merging. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=W70w5JCzdq. arXiv:2508.01148.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03099.

Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. MetaGPT: Merging Large Language Models Using Model Exclusive Task Arithmetic. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. doi: 10.18653/v1/2024.emnlp-main.102. URL https://aclanthology.org/2024.emnlp-main.102/. arXiv:2406.11385.

A DATASETS AND SPLITS

Our data sources are ISIC 2017 (Codella et al., 2018), CheXpert (Irvin et al., 2019), and PatchCamelyon (Veeling et al., 2018), and the six-task and held-out evaluations also use HAM10000 (Tschandl et al., 2018) and the RetinaMNIST and PathMNIST subsets of MedMNIST v2 (Yang et al., 2023).

The pretrained encoders are CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021), DINOv3 (Siméoni et al., 2025), RAD-DINO (Pérez-García et al., 2025), DINOv2 (Oquab et al., 2024), MAE (He et al., 2022), and BEiT (Bao et al., 2022).

Table 4 summarizes the three tasks comprising our controlled evaluation suite. Each task is formulated as a binary detection problem evaluated via AUROC using a single-logit classification head. Data splits are seed-deterministic and grouped by patient or lesion identifier where applicable, ensuring no identity leakage across split boundaries. Validation splits govern early stopping and hyperparameter tuning, while test sets remain strictly held out until final evaluation. To decouple domain semantics from dataset scale, every specialist is trained on a standardized $2{,}000$-image subset regardless of the underlying dataset size.

Table 4: Overview of the controlled binary trio. Positive class rates reflect the overall training pool, from which each specialist is trained on a uniform $2{,}000$-image subsample.

Task	Source	Train pool	Val	Test	Positive rate
Melanoma detection	ISIC 2017 dermoscopy	2,200	275	275	18.6%
Effusion detection	CheXpert radiography	51,787	6,473	6,474	32.7%
Tumor detection	PatchCamelyon histopathology	262,144	32,768	32,768	50.0%

B HYPERPARAMETER SEARCH SPACES AND GRID SENSITIVITY

Every tuned method undergoes exhaustive grid search on the validation set, selecting the configuration that maximizes mean validation performance across all three tasks. Methods with tunable scaling share the same coefficient grid.

To measure the impact of search space constraints, we evaluated performance across three expanding scaling ceilings, standard $\alpha \in (0, 1.0]$, intermediate $\alpha \in (0, 2.5]$, and extended $\alpha \in (0, 4.0]$. As highlighted in §4, standard search bounds ($\alpha \leq 1.0$) severely understate the performance of methods like LiNeS (0.763 vs. 0.811 AUROC) whose optimal scaling parameters consistently exceed unity. Expanding the ceiling to $\alpha_{\max} = 4.0$ yields full convergence across nearly all configurations, with boundary selection virtually disappearing across all cells. Table 5 details the final search grids and total trial counts.

Table 5: Per-method hyperparameter grids and trial counts. α denotes the shared scaling-coefficient grid $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.2, 1.5, 2.0, 2.5, 3.0, 4.0\}$.

Method	Grid	Trials
Simple Averaging	none	1
PCB-Merging	none	1
Task Arithmetic	α	11
Iso-C, Iso-CTS, TSV-Merge	α	11
TIES	$\alpha \times$ trim fraction $\in \{0.1, 0.2, 0.3\}$	33
DARE	$\alpha \times$ drop rate $\in \{0.3, 0.5, 0.7\}$	33
DARE-TIES	$\alpha \times$ drop rate $\times$ trim fraction	99
LiNeS	$\alpha \times \alpha_{\mathrm{L}} \in \{0, 0.1, 0.3\} \times \beta_{\mathrm{L}} \in \{0.5, 0.9\}$	66
Gram-LS	$\alpha \times \lambda \in \{0, 0.05, 0.1, 0.3\}$	44

Iso-CTS fixes the retained rank fraction at $0.5 \cdot \min(m, n)$ per weight tensor rather than treating it as a tuned hyperparameter. For Gram-LS, the regularization parameter λ is defined relative to the mean diagonal of G. The resulting scale-invariance enables a single hyperparameter grid to generalize across layers and backbones whose task-vector norms differ by more than an order of magnitude (Table 6). The zero-tuning strategies evaluated in Table 3 omit hyperparameter search entirely.

C SIX-TASK EVALUATION

We evaluate whether Iso-C continues to underperform tuned summation on a larger task set. The six-task suite adds HAM10000 melanoma detection, CheXpert cardiomegaly detection, and binarized PathMNIST tumor classification to the three primary tasks. This evaluation uses seven backbones and seed 42, whereas the primary benchmark uses three seeds.

Iso-C reaches 0.686 mean AUROC and tuned Task Arithmetic reaches 0.775, a gap of approximately 0.09. In the primary three-task benchmark, the corresponding means are 0.681 and 0.808, a gap of approximately 0.13. The narrower gap primarily reflects lower performance of tuned summation (-0.033 AUROC), with a smaller increase for Iso-C ($+0.005$). Iso-C therefore remains below tuned summation on the larger suite. Because task composition and seed coverage change alongside task count, this comparison does not isolate the effect of task count or explain the isotropization deficit.

D TASK-VECTOR NORM ANALYSIS

Table 6 reports the per-task task-vector ℓ_2 norms across evaluated backbones on the controlled binary trio (seed 42). Update norms remain remarkably uniform across tasks within a given backbone (column means range narrowly from 6.1 to 6.7), yet vary by roughly $30\times$ across backbone architectures. This structural variation includes CLIP's small task vectors, which coincide with its $5\times$ lower fine-tuning learning rate, contrasted with the larger update magnitudes of standard ViT-B models. DINOv2 is the one backbone whose norms are not uniform across tasks, with an effusion task vector of 0.97 against 3.10 and 3.58 for its other two tasks, and it is also the backbone with the largest across-seed spread in Appendix E. When task formulation and training budgets are strictly controlled, residual norm heterogeneity may reflect both backbone and optimization differences, and this comparison does not isolate their causal contributions.

Table 6: Task-vector ℓ_2 norms per backbone and task (seed 42).

Backbone	Melanoma	Effusion	Tumor
CLIP	0.63	0.84	0.84
ViT-B	21.93	22.28	22.35
DINOv3	3.25	3.05	3.59
RAD-DINO	5.45	5.37	5.51
DINOv2	3.10	0.97	3.58
ViT-MAE	6.10	5.59	5.64
BEiT	4.38	4.89	5.28
mean	6.41	6.14	6.68

E PER-BACKBONE RESULTS

Table 1 marginalizes over backbones, which hides the fact that backbone identity is the largest single source of variation in the benchmark. Tables 7 through 10 therefore report every (method, backbone) combination, first as the mean over the three tasks and then task by task. Each entry is the mean over the three seeds with the across-seed sample standard deviation, so the spread shown here is seed noise alone at a fixed backbone.

Seed noise is small for the tuned methods on the stronger backbones, falling below 0.015 AUROC for CLIP, ViT-B, and DINOv3, so the wide marginal intervals of Table 1 are driven by the backbone axis rather than by training stochasticity. DINOv2 is the outlier throughout, with an across-seed standard deviation reaching 0.194 for PCB-Merging on tumor detection and 0.177 on effusion, and it accounts for most of the marginal spread of the weaker methods. The ordering of the leading cluster also changes from backbone to backbone, with LiNeS best on CLIP, Gram-LS best on ViT-B and BEiT, and TSV-Merge best on ViT-MAE, with the aggregate paired tests not detecting differences among several leading methods. Iso-C and Iso-CTS trail the tuned scaled sum at every one of the seven backbones, so the isotropization deficit of §4 is not an artifact of one architecture.

Table 7: Per-backbone mean test AUROC over the three tasks (mean ± sample SD across 3 seeds).

Method	CLIP	ViT-B	DINOv3	RAD-DINO	DINOv2	ViT-MAE	BEiT
Oracle router	*0.885* ±0.005	*0.871* ±0.009	*0.870* ±0.011	*0.845* ±0.006	*0.812* ±0.004	*0.839* ±0.004	*0.852* ±0.019
Simple Averaging	0.817 ±0.013	0.799 ±0.025	0.784 ±0.004	0.747 ±0.016	0.670 ±0.111	0.762 ±0.016	0.774 ±0.019
Task Arithmetic	0.850 ±0.003	0.854 ±0.005	0.833 ±0.013	0.794 ±0.024	0.727 ±0.057	0.784 ±0.010	0.811 ±0.013
TIES	0.847 ±0.005	0.848 ±0.012	0.824 ±0.016	0.787 ±0.035	0.674 ±0.039	0.785 ±0.009	0.807 ±0.014
DARE	0.850 ±0.003	0.854 ±0.005	0.833 ±0.014	0.793 ±0.024	0.727 ±0.058	0.790 ±0.011	0.812 ±0.014
DARE-TIES	0.849 ±0.004	0.851 ±0.007	0.828 ±0.013	0.790 ±0.034	0.703 ±0.065	0.783 ±0.011	0.809 ±0.015
PCB-Merging	0.849 ±0.004	0.842 ±0.013	0.823 ±0.009	0.793 ±0.016	0.648 ±0.127	0.784 ±0.010	0.806 ±0.016
LiNeS	0.856 ±0.001	0.854 ±0.004	0.832 ±0.012	0.797 ±0.025	0.741 ±0.037	0.788 ±0.006	0.811 ±0.014
Iso-C	0.717 ±0.014	0.688 ±0.025	0.741 ±0.017	0.653 ±0.013	0.636 ±0.088	0.666 ±0.032	0.667 ±0.084
Iso-CTS	0.745 ±0.004	0.714 ±0.033	0.762 ±0.013	0.680 ±0.023	0.658 ±0.095	0.692 ±0.041	0.681 ±0.076
TSV-Merge	0.855 ±0.004	0.840 ±0.009	0.825 ±0.014	0.782 ±0.030	0.706 ±0.072	0.796 ±0.008	0.807 ±0.017
Gram-LS (ours)	0.851 ±0.003	0.856 ±0.004	0.831 ±0.012	0.796 ±0.032	0.717 ±0.063	0.790 ±0.014	0.813 ±0.017

Table 8: Per-backbone test AUROC on melanoma detection (ISIC 2017), mean ± sample SD across 3 seeds.

Method	CLIP	ViT-B	DINOv3	RAD-DINO	DINOv2	ViT-MAE	BEiT
Oracle router	*0.861* ±0.017	*0.850* ±0.022	*0.822* ±0.032	*0.739* ±0.019	*0.704* ±0.064	*0.822* ±0.014	*0.832* ±0.023
Simple Averaging	0.776 ±0.006	0.750 ±0.044	0.713 ±0.028	0.690 ±0.007	0.627 ±0.058	0.714 ±0.040	0.748 ±0.028
Task Arithmetic	0.807 ±0.009	0.840 ±0.016	0.772 ±0.048	0.698 ±0.041	0.619 ±0.052	0.757 ±0.019	0.775 ±0.023
TIES	0.793 ±0.023	0.813 ±0.042	0.747 ±0.063	0.707 ±0.048	0.570 ±0.024	0.759 ±0.032	0.769 ±0.025
DARE	0.807 ±0.009	0.839 ±0.017	0.773 ±0.049	0.697 ±0.040	0.620 ±0.054	0.762 ±0.027	0.775 ±0.023
DARE-TIES	0.802 ±0.017	0.829 ±0.019	0.764 ±0.056	0.702 ±0.053	0.576 ±0.079	0.752 ±0.026	0.774 ±0.028
PCB-Merging	0.808 ±0.006	0.818 ±0.027	0.759 ±0.036	0.726 ±0.017	0.661 ±0.045	0.748 ±0.030	0.780 ±0.019
LiNeS	0.815 ±0.009	0.839 ±0.015	0.773 ±0.046	0.705 ±0.047	0.620 ±0.052	0.762 ±0.014	0.772 ±0.028
Iso-C	0.680 ±0.037	0.657 ±0.044	0.639 ±0.081	0.577 ±0.031	0.579 ±0.055	0.645 ±0.015	0.647 ±0.099
Iso-CTS	0.715 ±0.029	0.702 ±0.038	0.664 ±0.060	0.595 ±0.022	0.604 ±0.073	0.659 ±0.010	0.666 ±0.089
TSV-Merge	0.820 ±0.016	0.836 ±0.017	0.766 ±0.019	0.680 ±0.033	0.616 ±0.077	0.761 ±0.022	0.777 ±0.030
Gram-LS (ours)	0.807 ±0.010	0.837 ±0.017	0.769 ±0.046	0.706 ±0.049	0.585 ±0.079	0.757 ±0.035	0.779 ±0.028

Table 9: Per-backbone test AUROC on pleural-effusion detection (CheXpert), mean ± sample SD across 3 seeds.

Method	CLIP	ViT-B	DINOv3	RAD-DINO	DINOv2	ViT-MAE	BEiT
Oracle router	*0.837* ±0.006	*0.814* ±0.005	*0.829* ±0.003	*0.863* ±0.003	*0.771* ±0.060	*0.772* ±0.002	*0.764* ±0.037
Simple Averaging	0.770 ±0.017	0.746 ±0.024	0.734 ±0.009	0.698 ±0.051	0.553 ±0.153	0.694 ±0.011	0.686 ±0.026
Task Arithmetic	0.807 ±0.004	0.785 ±0.008	0.781 ±0.011	0.814 ±0.026	0.704 ±0.045	0.709 ±0.005	0.725 ±0.040
TIES	0.811 ±0.006	0.791 ±0.002	0.777 ±0.014	0.803 ±0.044	0.593 ±0.099	0.706 ±0.008	0.715 ±0.052
DARE	0.808 ±0.004	0.785 ±0.008	0.781 ±0.012	0.814 ±0.027	0.704 ±0.045	0.714 ±0.013	0.725 ±0.040
DARE-TIES	0.810 ±0.004	0.785 ±0.009	0.772 ±0.026	0.806 ±0.031	0.641 ±0.123	0.708 ±0.011	0.717 ±0.047
PCB-Merging	0.805 ±0.005	0.779 ±0.014	0.772 ±0.007	0.763 ±0.049	0.499 ±0.177	0.705 ±0.009	0.718 ±0.036
LiNeS	0.810 ±0.004	0.785 ±0.007	0.777 ±0.018	0.811 ±0.019	0.707 ±0.025	0.712 ±0.014	0.726 ±0.041
Iso-C	0.662 ±0.016	0.576 ±0.019	0.697 ±0.019	0.612 ±0.102	0.620 ±0.049	0.574 ±0.074	0.634 ±0.020
Iso-CTS	0.689 ±0.022	0.610 ±0.053	0.716 ±0.033	0.644 ±0.107	0.633 ±0.065	0.604 ±0.082	0.643 ±0.021
TSV-Merge	0.807 ±0.004	0.764 ±0.017	0.760 ±0.047	0.784 ±0.046	0.636 ±0.125	0.730 ±0.003	0.724 ±0.035
Gram-LS (ours)	0.808 ±0.003	0.792 ±0.003	0.776 ±0.017	0.811 ±0.032	0.658 ±0.116	0.716 ±0.007	0.726 ±0.041

Table 10: Per-backbone test AUROC on tumor detection (PatchCamelyon), mean ± sample SD across 3 seeds.

Method	CLIP	ViT-B	DINOv3	RAD-DINO	DINOv2	ViT-MAE	BEiT
Oracle router	*0.956* ±0.002	*0.950* ±0.006	*0.959* ±0.001	*0.934* ±0.005	*0.962* ±0.001	*0.922* ±0.005	*0.959* ±0.002
Simple Averaging	0.904 ±0.020	0.901 ±0.018	0.906 ±0.016	0.854 ±0.009	0.830 ±0.190	0.877 ±0.002	0.887 ±0.007
Task Arithmetic	0.936 ±0.004	0.937 ±0.010	0.947 ±0.005	0.869 ±0.010	0.857 ±0.093	0.888 ±0.016	0.934 ±0.010
TIES	0.936 ±0.006	0.941 ±0.007	0.949 ±0.006	0.852 ±0.021	0.859 ±0.121	0.891 ±0.011	0.936 ±0.009
DARE	0.936 ±0.004	0.937 ±0.010	0.946 ±0.005	0.869 ±0.009	0.857 ±0.094	0.895 ±0.008	0.934 ±0.010
DARE-TIES	0.936 ±0.004	0.940 ±0.008	0.948 ±0.006	0.863 ±0.017	0.892 ±0.053	0.890 ±0.007	0.937 ±0.009
PCB-Merging	0.933 ±0.006	0.929 ±0.005	0.937 ±0.005	0.889 ±0.007	0.784 ±0.194	0.898 ±0.004	0.921 ±0.003
LiNeS	0.941 ±0.007	0.938 ±0.009	0.947 ±0.005	0.876 ±0.010	0.895 ±0.042	0.892 ±0.010	0.935 ±0.011
Iso-C	0.809 ±0.056	0.831 ±0.031	0.887 ±0.022	0.771 ±0.054	0.708 ±0.174	0.777 ±0.035	0.720 ±0.136
Iso-CTS	0.831 ±0.049	0.829 ±0.044	0.907 ±0.017	0.799 ±0.036	0.736 ±0.163	0.815 ±0.033	0.733 ±0.125
TSV-Merge	0.937 ±0.005	0.920 ±0.013	0.948 ±0.003	0.880 ±0.018	0.866 ±0.075	0.898 ±0.003	0.919 ±0.004
Gram-LS (ours)	0.937 ±0.005	0.939 ±0.007	0.947 ±0.005	0.871 ±0.013	0.909 ±0.041	0.896 ±0.005	0.933 ±0.010